

A call to arms

Biologists should involve themselves in the debate over biological weapons — both to ensure that we have the means to counter the threats that such weapons pose and to help keep those threats in perspective.

Over the past few weeks in Geneva, international negotiators have been wrestling with a text that aims to give teeth to the 1972 Biological Weapons Convention (BWC). Things are not going well. This draft protocol would set up a monitoring regime to enforce the BWC's ban on bioweapons development and production. But many observers see the current text as unworkable, and it seems certain that the United States — with the quiet backing of several other nations — will reject it.

Negotiations on the protocol have been rumbling on since the mid-1990s, spurred by revelations that Russia and Iraq — both signatories to the BWC — had operated biological weapons programmes. The deadline for the talks is a BWC review conference, planned for November. But the draft is being attacked from all sides.

Pharmaceutical and biotechnology companies object to provisions that would allow random inspections of their facilities, fearing that commercial secrets would be laid bare. Proponents of stringent arms control are also unhappy; any facility receiving a random visit would be given two weeks' notice, so those involved in biowarfare could easily disguise their activities. Even 'challenge' inspections, made following specific allegations that the convention had been breached, would allow a lag of 108 hours.

The likelihood that the BWC review conference will end in failure should concern us all. The US government takes the threat of a biological attack seriously, having placed an order worth more than \$300 million for 40 million doses of smallpox vaccine. It is also pouring larger sums of money into research on countermeasures. Meanwhile, some experts are warning that the threats posed by biowarfare will increase as our growing knowledge of microbial genomics and advances in techniques of genetic engineering create new possibilities for designing 'enhanced' biological weapons (see page 232).

Yet most mainstream biologists have had little engagement with these issues. Prominent researchers such as Matthew Meselson of Harvard University and Stanford University's Steven Block are doing their best to stimulate debate. Some scientific bodies are also playing their

part. But the failing negotiations in Geneva are not a hot topic for lab-bench gossip. In interviews for our article on engineered bioweapons, some researchers professed not to have given much thought to the potential for their work to be abused in this way. Others clearly had mulled it over, but were reluctant to discuss the issues publicly.

The first reaction may simply be naive, but the second is misguided. It is easy to understand why biologists are wary of having their work discussed in the context of warfare and terrorism — public suspicions of biotechnology are readily provoked. But the agencies charged with defending us against biological attacks are more likely to design effective countermeasures if the wider biological community starts thinking about the problem, and contributing its ideas.

By becoming more aware of the issues and engaging more vigorously in discussions on bioweapons, biologists can also help to ensure that threats are not blown out of proportion. The prospect of a crazed individual releasing anthrax in a major city is certainly a terrifying prospect. It is not, however, the most realistic terrorist threat. Developing a way to deliver an aerosol of anthrax into the New York subway poses technological challenges that are probably beyond the means of most terrorist groups — who in any case find that guns and bombs are usually adequate for their purposes.

Similarly, although engineered bioweapons are a real danger for the future, informed scientific debate helps put those risks in perspective. The defence industry has a tendency to meet military threats with expensive proposals that may be of dubious technical merit. Biologists have a duty to urge the development of biodefences that will save lives without wasting billions of dollars.

Of course, there are no simple answers. Doctors take the Hippocratic oath, but that did not prevent the atrocities perpetrated in the name of medicine under the Nazi regime. Neither did the arms-control activism of prominent physicists block the development of the hydrogen bomb. But if biologists stick their heads in the sand and pretend that their work has nothing to do with warfare, they will be doing the world a disservice. ■

Nature's world view

Introducing a survey of unprecedented scope.

This month sees the publication of *The Nature Yearbook of Science and Technology 2001*. *Nature* has published occasional surveys of science in places as varied as Russia and Mexico, and India and California, but never such a synoptic view of the entire world of research. Weighing in at 2,124 pages, the yearbook is an ambitious undertaking to systematically profile in detail the major scientific organizations in every country in the world.

Editorial content does not stop there. The volume also provides authoritative country analyses written by the Nature Publishing Group's network of correspondents worldwide, as well as by local journalists. Leading scientists, politicians and economists analyse in depth a selection of key contemporary issues, such as genomics and life sciences, neglected killer diseases, the impact of the Internet on the

practice of science, and efforts to curb weapons of mass destruction.

Anyone who has ever compiled a reference work involving information from diverse organizations worldwide will appreciate that the logistics of creating this first edition from scratch have been massive. In the next phase, feedback from those bodies and from readers worldwide will be incorporated to 'debug' this first volume. (To supply such offerings and to learn more, see http://www.macmillan-reference.co.uk/Science/Nature_yearbook.htm.) Closing gaps, correcting errors and extensive further editing will provide a second and further editions that will be even richer, more analytical and, hopefully, complete (whatever the latter might mean in such a context).

Scientists, policymakers, journalists and others: welcome to the beginning of *Nature's* world of science and technology. ■





Scaled up
Columbia River sees
big jumps in salmon
numbers
p226



In the dark
Power cuts threaten
research projects in
California
p227



Satellite plans
Taiwan gets COSMIC
over weather
forecasting
p228



Web wonder
Internet pioneer
becomes fellow of
the Royal Society
p230

Espionage charges threaten to undermine research relations

**Laura Bonetta, Washington
and David Cyranoski, Tokyo**

Two Japanese molecular biologists have been charged with espionage in the United States in a move that researchers in Japan say could sour scientific relations between the two countries.

The biologists are accused of stealing materials from the Cleveland Clinic Foundation (CCF) in Ohio and passing them on to the Institute of Physical and Chemical Research (RIKEN) in Japan, which is mainly funded by the Japanese government.

Researchers and government officials in Japan greeted news of the charges with some consternation. At RIKEN, where one of the biologists now works, officials denied that the centre had any involvement in espionage.

Japanese researchers are worried that the incident could make them targets for unwarranted suspicion in American laboratories. "Using words like 'espionage' could ruin any hope of establishing trust between researchers," says Akiyoshi Wada, director of RIKEN's Genomic Sciences Center. "It's scary."

Pharmaceutical industry officials are also concerned that the case could tarnish the



Serizawa (right) and Okamoto are charged with supplying material to RIKEN's Brain Science Institute.

reputation of one of Japan's premier research institutes, which has industrial contracts with many companies.

At a press conference on 11 May, the day after the charges were made public, a RIKEN official expressed his frustration that US

prosecutors could allege that RIKEN stood to benefit from the theft without first contacting the institute.

Takashi Okamoto and Hiroaki Serizawa were indicted by a federal grand jury last week on charges of economic espionage for

United States treads its own path on climate change

Irwin Goodwin, Washington

The US National Academy of Sciences has agreed to undertake a study of climate change for the Bush administration. A panel of scientists has pledged to complete the task in time for the next international climate-change meeting in Bonn on 17 July.

The panel includes some researchers who have criticized the International Panel on Climate Change (IPCC), which provides such advice on an international basis. One of the questions the US panel will consider is whether the IPCC's report summaries have accurately reflected their technical content.

Some observers fear that the Bush administration will use the study as a cover for its recent decision to abandon the Kyoto

Protocol, and to highlight differences between US and non-US scientists on the issue of climate change.

"It is apparent the White House is trying to develop a policy of its own, based on the review by American scientists, not on the consensus of the IPCC scientists," says Eileen Claussen, head of the Pew Centre on Global Climate Change in Arlington, Virginia, a think-tank that backs market-based approaches to reducing carbon emissions. "Having jumped off the cliff on climate change, the White House is trying to find a way for a soft landing."

Robert May, president of Britain's Royal Society, says that Bruce Alberts, president of the US National Academy, declined to sign a

soon-to-be-published interacademy statement on climate change because doing so would prejudice the US academy's new study. "I was pleased nevertheless that the National Academy was reviewing the science," says May. "It would be amazing and slightly alarming if they brought forth some conclusions that are not already known."

The National Academy of Sciences panel will be chaired by Ralph Cicerone, chancellor of the University of California, Irvine. It includes defenders of the IPCC such as Sherwood Rowland, a Nobel prize-winning atmospheric chemist, also at Irvine, and critics of it, such as Richard Lindzen, a meteorologist at the Massachusetts Institute of Technology.

▶ allegedly stealing DNA constructs and cell lines from Okamoto's laboratory in the CCF's Lerner Research Institute.

This is thought to be the first time that molecular biologists have been charged under the 1996 Economic Espionage Act, which chiefly aims to deter theft of trade secrets in areas such as computing. It is also the first of about two dozen prosecutions under the act in which the theft is alleged to benefit a foreign government.

Okamoto, who researches Alzheimer's disease, is alleged to have taken reagents from the laboratory that he headed for two and a half years shortly before moving to RIKEN's Brain Science Institute. He is also accused of sabotaging reagents and laboratory notes. RIKEN is carrying out its own internal investigation, which Okamoto says will clear him of the charges.

Serizawa, now at the University of Kansas Medical Center, was arrested by the Federal Bureau of Investigation (FBI) on 9 May. He was charged with helping Okamoto take the samples to Japan and with giving false information to the FBI. Two other unnamed accomplices were not indicted.

The indictment charges that Okamoto and an unnamed co-conspirator entered Okamoto's lab on the evening of 8 July 1999, removed some DNA and cell lines and destroyed the rest. The next day, the missing reagents were noted and the police, and then the FBI, were informed.

"We thought that the police should be called in," says George Stark, director of the Lerner Institute. "There was no question about what to do." But some researchers expressed surprise at the FBI's involvement — especially given the frequency with which biologists take materials when they change jobs.

On 26 July, after the investigation into the case had started, Okamoto resigned his position at the CCF to start his new post at RIKEN on 3 August. According to the indictment, Okamoto had accepted the job at RIKEN in April 1999. The indictment also says that Okamoto had shipped the samples to Serizawa in Kansas City and then retrieved them during a short trip in mid-August.

Okamoto and Serizawa are alleged to have committed economic espionage by stealing "trade secrets" from the CCF, "specifically, 10 DNA and cell-line reagents", to benefit RIKEN.

The charge that the Japanese government benefited from an economic espionage crime is a politically sensitive one — especially as the United States will need Japan's cooperation to extradite Okamoto.

Economic espionage charges carry a stiff penalty of up to 15 years in prison, \$500,000 in fines, or both. Okamoto and Serizawa are also charged with transport of stolen property across state lines and conspiracy. ■



On the up: salmon populations have reached record levels in the rivers of the US Pacific northwest.

Record salmon populations disguise uncertain future

Julie Wakefield, Washington

Salmon are swimming up streams in the Pacific northwest from the ocean in record numbers this spring, raising questions about the measures needed to conserve them.

So far, more than 345,000 spring and summer chinook salmon (*Oncorhynchus tshawytscha*) have passed the lowermost Bonneville Dam on the Columbia River. And more than 115,000 of these have already hurled themselves past eight major dams and reached the Snake River to spawn in fresh waters — a 25-fold increase on the average over the past 10 years.

Survival rates are also up. Scientists conservatively estimate that more than 21,000 wild chinook salmon, roughly 1.5% of recent years' smolt, will reach spawning grounds this season. This is more than double last year's figure of 8,900 and over 18 times the record low of 1,100 in 1995, according to fisheries biologist Charlie Petrovsky at the Idaho Department of Fish and Game.

Of the millions of smolt released from hatcheries in recent years to supplement the natural population, almost 1% survived the 1,000-kilometre journey. Survival rates in the past decade had been as low as 0.04%.

This year's salmon run is the largest on the Columbia River Basin since records began in 1938, and the best for wild chinook since the late 1970s. It supports theories that ocean climate cycles are a key influence on salmon populations, researchers say.

"The increase in run size is due to increased survival in the ocean," says Ted Bjornn, a University of Idaho researcher who monitors runs.

Chinook are the most studied of the 26 salmon species listed under the US Endangered Species Act, and are considered a good

barometer of salmon health in the Columbia River Basin, once the world's chinook capital.

But ecologists remain cautious. "These high returns should not cause us to assume the crisis is over and salmon are back," says Peter Kareiva, a conservation biologist at the National Marine Fisheries Service's Northwest Fisheries Science Center in Seattle. He says they reflect a period, beginning in 1999, of cooler waters, more food and fewer warm-water predators. Periods of warmer surface temperatures would, he says, put the pressure back on the salmon population.

Hydroelectric development, mining, overfishing, irrigation and urbanization have all been blamed for declining populations of wild salmon in the region. But with the population apparently surging — and electrical power growing short in the region (see opposite) — demands by environmentalists for dams on the Snake River to be demolished to conserve the salmon may lose momentum.

The new data are unlikely to quell controversy over management policies for the dams and rivers, however. In early May a coalition of environmental groups filed a suit against the National Marine Fisheries Service, alleging that its management plan for the salmon violates the Endangered Species Act. Environmentalists also claim that, under current policies, wild spring and summer chinook on the Snake River will be extinct by 2016.

But Kareiva says the picture is far more complex than this. "Our analyses point to serious extinction risks, but also to considerable opportunities for recovery if improvements are made in several different risk factors — hatcheries, habitat, harvest and hydropower." Researchers estimate that full recovery of wild salmon will require a smolt survival rate of between 2 and 6%. ■

Californian labs feel the heat of energy crisis

Rex Dalton, San Diego

Power-hungry research laboratories in California face their first blackouts this summer, after the withdrawal of exemptions that protected their electricity supplies.

Laboratories operating for the Department of Energy and NASA have been told that their power may be shut off for up to two hours at a time. Most electricity users in the state have faced similar cuts this month as warmer weather has increased the use of air conditioning. The blackouts are expected to become more frequent over the summer.

California has faced growing electricity shortages after a botched attempt to deregulate the state's electricity industry.

The Lawrence Berkeley laboratory and the Stanford Linear Accelerator Center (SLAC) have been told to expect notices from the California Public Utilities Commission (PUC) saying they will no longer be exempt.

The power cuts are likely to play havoc with long-planned experimental schedules. Cuts without warning may damage expensive equipment and could even endanger staff, say officials, who are appealing to the state to restore their exempted status.

So far the PUC has not issued a decision. "It is a very chaotic situation," says particle physicist Pier Oddone, deputy director of the Lawrence Berkeley laboratory, where the equipment likely to be affected includes

the Advanced Light Source.

Officials at SLAC say that it has reduced its power use voluntarily, and that power cuts would wreck its main high-energy physics experiments because of the time they take to shut down and restart. "A blackout would be a total disaster," says acting deputy director Gregory Loew. "We are trying to figure a way to mitigate this."

The PUC has already sent warnings to NASA's Ames Research Center at Moffett Field, which operates energy-intensive wind tunnels, and to a test facility for high explosives operated by the Lawrence Livermore nuclear weapons laboratory. The main Livermore laboratory — like Berkeley and SLAC, a US Department of Energy (DOE) facility — is expected to stay exempt because of its nuclear weapons work, says DOE official Mark Clark. But officials fear that Livermore may face energy restrictions in the months ahead.

At the General Atomics fusion facility in San Diego, which has not been exempt from the power cuts, the experimental schedule has already been disrupted. The normal January-to-September schedule for operating the DIII-D tokamak device has been shortened to complete the necessary work by mid-June, after which the crisis is likely to worsen. The facility's electricity bill has jumped from \$1 million to \$2.5 million a

year, prompting staff cuts and delays in purchases, says physicist Ronald Stambaugh, director of the DIII-D programme.

The other laboratories avoided the blackouts that swept California last winter because they held electricity contracts with another division of the DOE. But the PUC issued an order last month to change the rules governing such contracts. The new rules grant exemption only to users such as hospitals, and water and sewage facilities. ■



Cutting edge: the power shortage is already affecting the DIII-D tokamak (top) and SLAC.

GENERAL ATOMICS

SLAC

Angry researchers pour scorn on astrology classes

K. S. Jayaraman, New Delhi

A proposal by the Indian government to encourage universities to teach astrology has sparked a storm of protest among scientists.

The University Grants Commission (UGC) has offered to fund fully fledged departments of astrology with five teaching posts, a library, computer laboratory and horoscope bank. To be called *Jyotir Vigyan*

('astrological science' in Sanskrit), the departments are to be set up for the 2001–2002 academic year. They will offer bachelors, masters and doctoral degrees.

The proposal is the brainchild of science minister Murli Manohar Joshi, who is also minister for education and a powerful figure in the ruling Bharatiya Janata Party. Joshi, a physicist, believes that all answers sought by scientists are buried in the ancient Sanskrit writings called *Vedas* and *Upanishads*.

Leading researchers have condemned the move as an attempt to legitimize pseudoscience and superstition, and some have said that it undermines India's scientific credibility. The National Science Academy has expressed strong opposition. But there is no sign of the government relenting.

Meanwhile, 35 of India's roughly 200 universities have sought permission to set up courses, with more expected to follow.

"At a time when research in fields of pure science is being affected for want of funds, there is no justification in spending huge amounts on pseudoscience called Vedic astrology," said Pushpa Bhargava, founding

director of the Centre for Cellular and Molecular Biology in Hyderabad.

Defending the move, UGC chairman Hari Gautam said that astrology qualifies as a science, which he defined as "a subject that needs probing, investigation and research".

Two prominent UGC members — S. K. Joshi, a physicist and former director general of the Council of Scientific and Industrial Research, and Sipra Guha-Mukherjee, a plant molecular biologist from Jawaharlal Nehru University in New Delhi — seem to have consented to the move. And some researchers, including Vijay Bhatkar, who developed India's first supercomputer, have publicly backed it.

"There is no doubt that the move is tantamount to giving [a] certain amount of formal recognition to astrology as a science," said Valangiman Ramamurthi, secretary to the Department of Science and Technology. "I do not want to comment whether it is right or wrong, but if anyone comes to me with a research proposal on astrology we will evaluate it to see if it makes sense scientifically before funding it." ■



Epidemiology gains an ally in bioweapons surveillance project

Jonathan Knight, San Francisco

A system designed to track attacks involving bioweapons could also monitor the worldwide spread of infectious disease, its developers say.

Al Zelicoff, a disease surveillance expert at Sandia National Laboratory in New Mexico, presented evidence supporting the idea to a meeting of disease surveillance experts at Stanford University in Palo Alto, California, on 11 May.

The US government funded the Rapid Syndrome Validation Project (RSVP) to monitor background levels of illness against which a bioterrorist attack would stand out. "If you don't know what the noise is in a community, you sure can't find a signal above it," Zelicoff says.

At the moment, monitoring relies on clinicians to report dangerous diseases, he adds. But often they fail to do so, because most see so few cases of the maladies that they do not recognize them. As a result, outbreaks can take several weeks or more for epidemiologists to spot (see News Feature, page 232).

RSVP, which has been in use at the University of New Mexico hospital in Albuquerque since November, does not require a diagnosis. Instead, the physician enters into a computer terminal the details of any patient with one of six syndromes, such as acute respiratory distress syndrome, influenza-like illness or acute hepatitis.

The information goes over the Internet to a server in Zelicoff's office, where software identifies combinations of symptoms likely to indicate a dangerous illness. These details are then forwarded to the New Mexico state epidemiologist.

This rapid reporting already appears to have this year averted two outbreaks of hepatitis A. In both cases, the physician ordered blood tests to determine the type of hepatitis only after prompting from a state official notified by RSVP.

The physicians using RSVP also prescribed far fewer antiviral drugs for flu this year than last, probably because they got immediate information on whether or not there was any flu in their patient's community.

A major private health organization in New Mexico plans to adopt RSVP, Zelicoff says, and he is also seeking funding to try the system in other countries, where any doctor with an Internet link could potentially join it. ■

Array system promises global atmospheric monitoring

David Cyranoski, Tokyo

A Taiwan-US collaboration is hoping that its constellation of microsatellites equipped with Global Positioning System (GPS) receivers will provide a valuable new approach to meteorology, climatology and research into 'space weather'.

Most weather forecasting currently relies on balloons that take readings such as temperature and humidity on their way up from some 900 locales worldwide. But these points are restricted to land, ruling out truly global weather models. Weather satellites give wider coverage, but they gather data by looking straight down to Earth, yielding little information about what is happening at various different altitudes.

But the Taiwan-led Constellation Observing System for Meteorology, Ionosphere and Climate (COSMIC) could change that by using an array of microsatellites and a novel technique to improve the coverage and accuracy of data collection.

The six COSMIC microsatellites, scheduled to be launched in mid-2005, will pick up radio signals from 28 existing GPS satellites as they pass through the Earth's atmosphere. The microsatellites will observe the refraction (or bending) of the signals, and infer information about atmospheric density from it, at all altitudes. From the density data, researchers will be able to deduce the pattern of pressure and temperature.

"We can also calculate atmospheric moisture near the surface, construct pressure contours, and derive wind fields and other critical quantities," says Tom Yunk of NASA's Jet Propulsion Lab (JPL), which did much of the early work on the technique.

The most important advantage is

coverage. "The microsatellite constellation will measure some 4,000 points spread uniformly around the globe, with high accuracy," says Chao-Han Liu, president of Taiwan's National Central University.

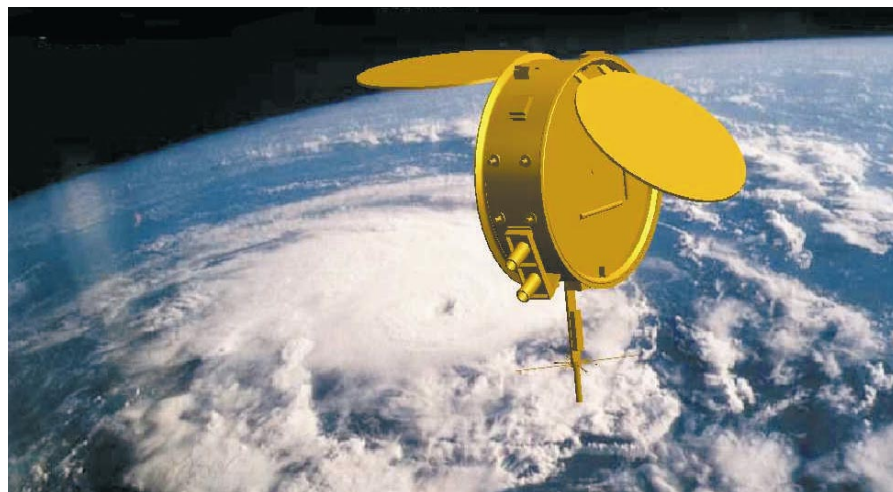
Researchers on 'space weather' are also excited about the project, says Liu. In the ionosphere, at altitudes of about 80 km, electron density can be measured in a similar fashion to the atmospheric density. This will provide valuable information for predicting magnetic storms, which can affect the operations of satellites and power grids.

The COSMIC system has the potential to significantly improve climatological measurements, says Alan Thomas, director of the Global Climate Observing System secretariat. However, he warns that it will probably take time to develop a reliable system that produces measurements for climate-change applications, and another 20 years or so after that to obtain a meaningful long-term climate data set.

The agreement to build the system, signed on 3 May, involves Taiwan's National Space Program Office (NSPO), JPL, the US University Corporation for Atmospheric Research in Boulder, Colorado, and several US universities.

Taiwan will pay US\$80 million of the estimated US\$100 million total project cost and will build the satellites with the help of Orbital Sciences Corporation of Dulles, Virginia, which made a prototype version of the satellite in 1995 for a 'proof-of-concept' experiment. Taiwan will also operate the mission.

"This is a chance to get people really interested in space science," enthuses Luo-Chang Lee, director of NSPO. ■



Saturation coverage: six microsatellites will improve the coverage and accuracy of climate data.

NSPO

Fresh funding offers lifeline to bioinformatics centre

Declan Butler, Paris and Alison Abbott, Munich

The European Bioinformatics Institute (EBI) based near Cambridge in Britain is to receive an infusion of 19 million euros (US\$16.6 million) from the European Commission over the next three years.

The cash injection marks an important turning point for the EBI. In 1999, the institute faced bankruptcy after the European council of research ministers decided that the fifth Framework funding programme would not support core funding and operational costs for infrastructure.

But after widespread protests, the European Commission effectively reversed this decision last November and earmarked 25 million euros over three years for "genomic and proteomic databases and repositories of suitable animal models".

Programmes coordinated by the EBI will receive the lion's share of this sum. Among the non-EBI projects to benefit is the European Mouse Mutant Archive (EMMA), centred in Rome, which will get 4 million euros. "This has been a real life-saver for us," says Martin Hrabé de Angelis, EMMA's scientific director.

Of the total grant to the EBI, 4.3 million euros each year will go directly to the institute for its own operations, adding to existing support of 8.4 million euros per year from the European Molecular Biology Labo-



On the up: the European Bioinformatics Institute is benefiting from an injection of cash.

ratory — set to increase to 11.4 million euros by 2005 — and annual funding of around 2 million euros from the Wellcome Trust.

A new plan to integrate biological databases, known as Integr8, is the largest single project to be supported with the funds. This will develop standards, databases and software to allow scientists to draw on relevant material from all biological databases and the scientific literature. "This is perhaps the project I'm most excited about," says Graham Cameron, joint head of the EBI.

In addition, 6 million euros will be spent implementing a planned European Macromolecular Structure Database. This is intended to serve as a counterpart to the US Protein Data Bank. ■

Physicists put a value that matters on the standard model

Josette Chen

Experimental physicists have derived a more definitive value for the parameter that explains the surplus of matter over antimatter when the Universe began. But the result has left theorists with some explaining to do.

At a 10 May seminar at CERN, the European Laboratory for Particle Physics in Geneva, the researchers announced that they had obtained measurements that further confirmed the existence of 'charge-parity (CP) violation' — the phenomenon predicted by the standard model of particle physics to be behind the existence of matter in the Universe.

The CERN team reported a value of $15.3 \pm 2.6 \times 10^{-4}$ for the parameter linked to CP violation. "These errors are considerably smaller than any reported before," says Frank Wilczek, a theoretical physicist from the Massachusetts Institute of Technology.

But theoretical calculations forecast a larger range of values for the parameter of between 4.2×10^{-4} and 13.7×10^{-4} (see *Nature* 402, 22–23; 1999). "It is hard to say whether these new results are consistent with the standard model," Wilczek says. "The torch has now been passed on to the theoretical side." ■

Resuscitated 'alien' microbes stir up an Italian storm

Alison Abbott

Two Naples scientists claim to have found extraterrestrial bacteria in ancient rocks and meteorites from one of the city's museums. Italy's national academy of sciences, the Accademia dei Lincei, agreed to publish the controversial result after a heated debate on 11 May.

Bruno D'Argenio, a geologist and academy member, and Giuseppe Geraci, a molecular biologist, told the academy's general meeting in Rome that they had brought back to life dried up bacteria, which they believe had been locked up in the rocks for millions of years.

The scientists say they did this by simply adding culture medium to samples of the rock and meteorite placed on glass slides. They argue that their techniques exclude the possibility of contamination, and claim that the results prove that life exists, and perhaps even evolved, in outer space.

But most biologists at the meeting were unconvinced that contamination could be

ruled out. And they said they were concerned that the authors had bypassed peer review by presenting their findings directly to the academy.

The sceptics were also angry that the scientists had announced their results at a

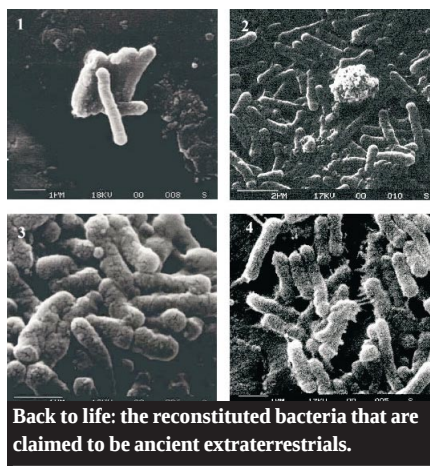
press conference two days before the academy meeting. D'Argenio and Geraci won widespread media coverage in Italy and abroad after they showed journalists a movie of the reconstituted bacteria swimming.

According to Mario Stefanini, a cell biologist at the University of Rome, academy rules allow members to present their data to the academy, and then to publish them as part of the academy's proceedings. The academy president, mathematician Edoardo Vesentini, ruled that in this case the paper should be accompanied by a transcript of the debate.

Russell Vreeland of West Chester University in Pennsylvania says experts in the field know how difficult it is to avoid contamination in such research, and that "this is why peer review is so important".

Tom Shepherd of the British Geological Survey at the University of Leicester concurs, saying of the results: "The community is cautious, and quite rightly so." ■

www.cryms.org



Back to life: the reconstituted bacteria that are claimed to be ancient extraterrestrials.

\$1.1bn gift boosts Stowers Institute's expansion plan

Washington The Stowers Institute for Medical Research, a developmental genetics research centre based in Kansas City, Missouri, has received a \$1.1 billion gift from the couple whose name it bears.

Last week's donation takes the centre's endowment, all of which has been provided by Jim and Virginia Stowers, to \$1.6 billion. Mr Stowers founded American Century Companies, a leading US investment management firm.

The new funds should allow the institute to hasten its recruitment efforts and meet its goal of having 16 functioning labs in place by the autumn of 2002, says the centre's president, William Neaves. Heads of four research groups are already in place, with five more to arrive this summer. Neaves hopes the centre will house 50 groups by 2010.

♦ <http://www.stowers-institute.org>

New management for troubled chimp facility

San Diego Management of the Alamogordo Primate Facility in New Mexico, which houses many of the chimpanzees used in US

research but has been repeatedly cited for animal-welfare violations, will be transferred to a new company, say officials at the National Institutes of Health (NIH).

Charles River Laboratories of Massachusetts, which specializes in providing animal models of human disease, will manage the facility for 10 years, taking over from the Coulston Foundation.

Research on the 250 chimpanzees at the facility has already ceased. NIH officials say that the new \$40 million management contract will ensure that the animals are properly cared for during their retirement. Until now, the health-research agency has struggled to find funding to support the animals once the research on them has ended.

Web inventor becomes Royal Society fellow

London In a move that acknowledges charges of elitism in its membership policies, Britain's Royal Society has given a fellowship to Tim Berners-Lee, who is credited with inventing the World-Wide Web. He is among 43 new fellows elected to the society this week.

Berners-Lee already has a Royal Medal from the society, but its new president Robert May expressed surprise earlier this year that none of the fellows had nominated him for a fellowship (see *Nature* **410**, 13; 2001). Now



Berners-Lee: a fellow of the Royal Society.

at the Massachusetts Institute of Technology, Berners-Lee played a key role in devising the address system for the World-Wide Web in the 1990s while at CERN, the European Laboratory for Particle Physics near Geneva.

Other new fellows reflect a trend towards a more broadly based fellowship. They include astronomer Patrick Moore and Richard Dawkins, who is currently professor of public understanding of science at the University of Oxford. Both men are known for their work in popularizing science.

Biomedical body attacks move to boycott journals

Washington The Federation of American Societies for Experimental Biology (FASEB), a coalition of US biomedical research societies, has denounced the Public Library of Science's call for a boycott of journals this autumn as "coercive".

More than 20,000 scientists from around the world have signed the online petition in

support of a boycott of journals who refuse to deposit their papers in a free-to-access online library within six months of publication (see *Nature* 410, 728–729; 2001). The federation, which has 60,000 members, argues that the proposals need serious discussion, but that the petition does not encourage this.

FASEB says that its 21 member societies should be able to determine their policies for online access in a way that will ensure that their publications continue to be viable.

► <http://www.faseb.org>

French centre to study transgenic mice

Paris A new centre devoted to the study of transgenic mice is to be built at the Institute of Genetics and Molecular and Cellular Biology (IGBMC), near Strasbourg, France.

The centre will house between 60,000 and 70,000 mice, obtained through gene 'knock-outs' and other transgenic techniques. All aspects of the genetic modifications will be studied, from developmental effects to changes in the behaviour and physiology of adult mice. Researchers hope to use the results to develop new therapies for human disease.

Scientists expect the number of mice mutants to increase dramatically as the

sequencing of the mouse genome nears completion. A publicly funded consortium has unassembled data for 94% of the genome, while Celera Genomics of Rockville, Maryland, claims its subscription-only version is already complete (see *Nature* 411, 121; 2001).

Green Bank Telescope checks out Venus

Washington The world's largest steerable radiotelescope, the Green Bank Telescope in West Virginia, has made its first scheduled scientific observations.

Planetary scientist Donald Campbell from Cornell University in Ithaca, New York, and colleagues used an even larger fixed dish in Arecibo, Puerto Rico, to send radio signals

towards Venus. The Green Bank Telescope collected the signals after they had bounced off the planet's surface. The project is the first to cover large areas of the planet since NASA's Magellan mission in the early 1990s. The team used the same technique to



On the record: Green Bank scans the skies.

study an asteroid 150 metres in diameter, known as 2001 EC16.

The Green Bank Telescope began functioning last August, six years late and millions of dollars over budget (see *Nature* 406, 816; 2000).

► <http://www.gb.nrao.edu>

Italian politicians defend pasta source

Rome Italian politicians rushed to defend their national dish in the run-up to last week's general elections, in the mistaken belief that a German newspaper had accused the country of producing radioactive spaghetti.

The German daily *Frankfurter Allgemeine Zeitung* reported that the International Atomic Energy Agency in Vienna had registered 2,252 breeds of plant that had been genetically transformed by radiation-induced mutation in the 1970s and 1980s. A new variety of durum wheat, which is used to make pasta, was one of many commercially successful new breeds.

The newspaper article did not actually mention pasta, but it nonetheless provoked a robust response from the politicians. Outgoing agricultural minister Alfonso Pecoraro Scanio told the Italian newspaper *La repubblica* that "the German media is the thing that is genetically modified".

NRAO/AUT

The bugs of war

Could our knowledge of microbial genomics and skill in genetic engineering be used to create 'enhanced' bioweapons? Carina Dennis assesses the threat, and the efforts to counter it.

Ron Jackson and Ian Ramshaw weren't looking for trouble. Jackson, who works at the Pest Animal Control Cooperative Research Centre in Canberra, and Ramshaw, who is in the same city at the Australian National University, were searching for a way to control the mice that are serious pests in Australia. They wanted to make a contraceptive vaccine by altering the genes of the mousepox virus.

But in January the project gained notoriety after the pair inadvertently created an unusually virulent strain of mousepox. If a similar genetic manipulation were applied to smallpox, the scientists realized, this feared killer could be made even more dangerous. When they published their paper¹, it was only after much discussion about the



Present arms: anthrax, portrayed in this montage, is a formidable weapon. But could it be made worse?

wisdom of drawing attention to the findings. "It has to be brought out into the public arena so the situation can be addressed," argues Ramshaw.

The incident has heightened concerns about the potential for using genetic engineering to create biological weapons that surpass the destructive potential of natural pathogens. With the decoding of a pathogen's entire genome now commonplace, and transgenic techniques advancing all the time, some researchers believe that the sinister potential of biology can no longer be ignored.

Most experts feel that the hype surrounding bioengineered weapons still outweighs the threat, and argue that the main focus for concern should remain on conventional biological agents. But government agencies are

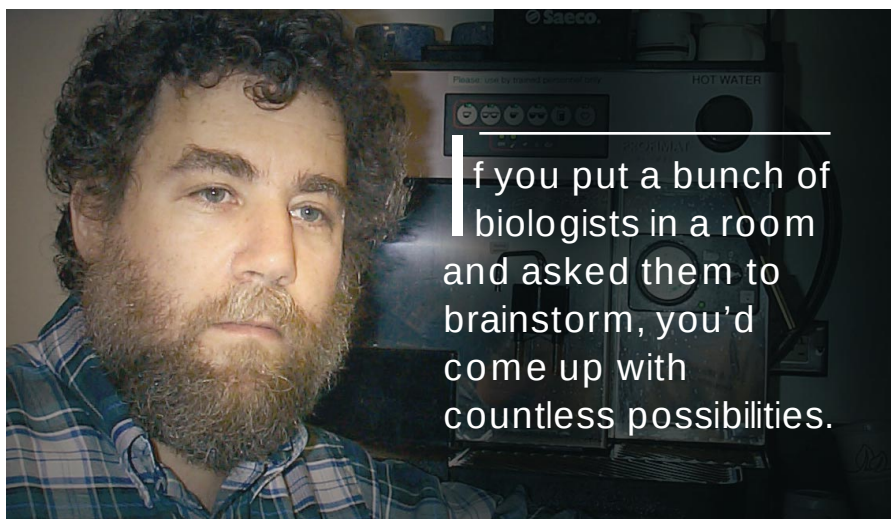
already working on methods of detecting disease outbreaks caused by genetically engineered organisms.

The potential for bioengineering agents of destruction was considered in 1997 by JASON, a group of mostly academic scientists that provides technical advice to the US government. In theory, it might be possible to build novel bacteria or viruses from a set of component parts — although most experts don't yet see that as a realistic scenario for creating bioweapons (see 'Starting from scratch', overleaf).

But making subtle genetic alterations to existing pathogens to increase their virulence or durability in the environment, or to make them harder to detect or to treat with drugs, is within the limits of today's technology². "If you put a bunch of biologists in a room and asked them to brainstorm, you'd come up with countless possibilities," says Steven Block, a biophysicist at Stanford University in California who led the JASON study.

Resistance is fertile

Perhaps the simplest way to 'enhance' a bacterial bioweapon is to make it resistant to antibiotics. Bacteria such as *Staphylococcus aureus* and *Pseudomonas aeruginosa* can evolve drug resistance at a startling rate. Some of the genes that convey this antibiotic resistance lie in the bacteria's genomes. But others can be carried on plasmids, circular pieces of DNA that replicate themselves independently — and the same is true of genes for other important traits such as virulence and infectivity. Plasmids can move between bacteria, and genes carried on plasmids can be incorporated into the



Steven Block believes some genetic technologies have serious implications for biowarfare.

CNRI/SPL/CORBIS

genome, which means that genes conferring antibiotic resistance, or any other advantageous trait, can spread rapidly.

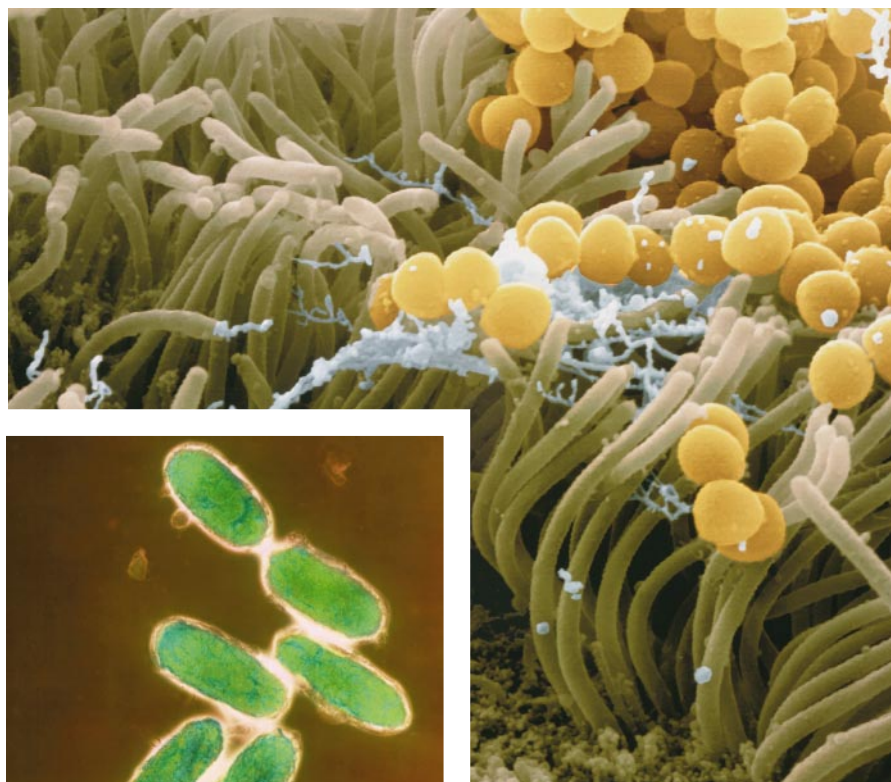
Molecular geneticists have long exploited plasmids as a means of cloning DNA and creating transgenic bacteria. So splicing genes for antibiotic resistance into plasmids and introducing them into a bacterium being developed as a biological weapon would, by the standards of today's top biology labs, be child's play. Anthrax, for instance, is typically treated using derivatives of penicillin, but they could be rendered ineffective by introducing a gene for the enzyme β -lactamase, which disables the antibiotics, into the anthrax pathogen, *Bacillus anthracis*.

According to Alastair Hay, an expert on biological warfare at the University of Leeds in the UK, manipulations of this type may already have been done. Hay helped to debrief defectors from Biopreparat, a clandestine network of facilities spread across Russia and Kazakhstan that worked on biological warfare until 1992. The scientists claimed that Biopreparat had developed a form of *Yersinia pestis*, the causal agent of plague, that was resistant to 16 different antibiotics.

Using the same techniques, it might be possible to transfer genes for pathogenicity. For example, the gene for the deadly toxin produced by the food-poisoning organism *Clostridium botulinum* could be introduced into ubiquitous bacteria such as *Escherichia coli*. More likely, says Block, bioengineers could take an existing biowarfare agent and add further virulence genes from other microorganisms.

Mix and match

Advances in genomics have greatly increased the possibilities of mixing and matching traits from different microorganisms. Among the complete pathogen genetic sequences now available are *P. aeruginosa*³, plus the bacteria responsible for tuberculosis⁴ and cholera⁵. This year's crop



Customized killers: in theory, genes for antibiotic resistance taken from *Staphylococcus aureus* (above) could be transferred to *Yersinia pestis* (left), the plague pathogen.

includes the leprosy bacterium⁶ and *E. coli* O157:H7 (ref. 7), a strain rendered deadly by the acquisition of a gene for a toxin that damages the kidneys. Only last month, a Japanese team published the genome sequences of two antibiotic-resistant strains of *S. aureus*⁸. The complete sequences of *B. anthracis* and *Y. pestis* are expected to follow later this year.

Some researchers are concerned about

potential abuses of DNA sequence data. But Tim Read, who leads the team sequencing *B. anthracis* at The Institute for Genomic Research in Rockville, Maryland, is confident that a policy of releasing the sequence information into publicly accessible databases is best. "My feeling is that releasing the data will tilt the scientific advantage towards bio-defence and force malevolent interests to work harder," he says. "The release of the data has stimulated research into vaccines, drugs and diagnostics."

One disturbing possibility is that knowledge of pathogen genomics could be combined with insights gleaned from human genetics to target particular ethnic groups. But most experts are sceptical of the potential for bioweapons as agents of ethnic cleansing. Although it is true that certain ethnic groups are unusually susceptible to particular pathogens, genetic variation in susceptibility to disease is likely to be greater within ethnic groups than between them. "You'd have a lot of collateral damage," predicts Paul Ewald, an expert on the evolution of disease at Amherst College in Massachusetts.

The accumulation of genomic and other biological data is also spawning a boom in computational biology, which uses mathematical modelling to help understand how networks of genes and proteins work. Although such approaches might yield information on drug targets, they could also highlight vulnerabilities that could be



Accidental architect: Ron Jackson co-engineered a particularly virulent form of mousepox.

JÜRGEN BERGER/SP

BARRY DOWSETT/SP

CSIRO

exploited by malevolent biologists.

But the information being generated by genome projects is not the only concern. Thanks to recent advances in biotechnology, would-be weapons bioengineers need not limit themselves to working with genetic sequences evolved through natural selection. Several companies are developing techniques of 'directed molecular evolution', which can be used to accelerate the evolution of desired traits by deliberately introducing genetic variation and then applying artificial selection.

Unnatural selection

One of the most powerful of these methods is DNA shuffling, developed by Willem Stemmer, chief scientist with the company Maxygen in Redwood City, California. Multiple copies of a given gene are first shattered into fragments, then reassembled using a variation of the polymerase chain reaction — a standard tool for copying sequences of DNA. This produces a range of 'daughter' genes with the fragments stitched together in subtly different ways. The enzymes involved in the reassembly process are also prone to errors, which introduce point mutations, further adding to the genetic diversity. The daughter genes can then be reintroduced into bacteria, which are selected to identify those with the desired traits^{9,10}.

Stemmer has also refined the technique to reassemble fragments taken from families of related genes from different bacteria¹¹. Most recently, his team started to shuffle



Protect and survive: emergency services in South Carolina practise decontamination techniques.

entire genomes in commercially valuable microorganisms. "We have recapitulated what could be accomplished in 15 years of classical recombination and selection in about six months," says Stemmer.

Maxygen is using the technique to develop better drugs and other proteins. But if it were to get into the wrong hands, Block considers the method to have "serious implications for biowarfare". Stemmer believes DNA shuffling is too sophisticated to be used by a lone bioterrorist. But it might not be beyond the capabilities of a bioweapons lab sponsored by a 'rogue' state.

Indeed, the potential of DNA shuffling as

a tool for bioweapons development was demonstrated in its first application. In this, Stemmer focused on a gene for β -lactamase, creating strains of *E. coli* that were 32,000 times less sensitive than wild-type bacteria to the antibiotic cefotaxime. Shortly after his paper was published⁹, Stemmer says he received a letter from the American Society for Microbiology expressing concerns about potential misuse and asking that he destroy the strain — which he did.

Designer destruction

Other approaches that might be used to develop bioweapons include the deliberate hybridization of related viral strains. Although most crosses of viruses are less potent than the parent strains, sometimes virulence increases — some virulent strains of flu, for instance, arise as naturally occurring recombinants between different influenza viruses¹².

Potential bioweapons designers might also be watching developments in gene therapy. Attempts to introduce therapeutic genes into patients' tissues rely mostly on weakened forms of various viruses. These vectors have yet to introduce genes efficiently and reliably. But if researchers can make them do so, similar vectors might also be used to ferry harmful genes into unsuspecting victims.

The experience of Jackson and Ramshaw, meanwhile, shows that scientists can stumble across possibilities for bioweapons quite by accident. To create their mouse contraceptive vaccine, they took a relatively benign strain of the mousepox virus, and added genes for proteins carried on the surface of mouse eggs. The idea was that cells infected by the viruses would churn out the proteins, causing female mice to produce antibodies against their own eggs. To maximize the vaccine's effectiveness, Jackson and Ramshaw also engineered the virus so that it contained the gene for interleukin-4 (IL-4), a protein

Starting from scratch

In January 1999, at the American Association for the Advancement of Science's annual meeting in Anaheim, California, genomics pioneer Craig Venter announced that scientists at his institute were contemplating creating novel bacteria. He claimed that they could chemically synthesize a 'minimal' genome and insert it into a bacterial cell stripped of its own DNA.

But Venter said the project had been put on hold, pending an ethical review. In addition to worries about scientists 'playing God', Venter cited fears that bioterrorists might adapt synthetic bacteria to make weapons that would evade conventional diagnostic tests.

The work on which Venter's claim was based was published later that year¹⁴. Researchers at

The Institute for Genomic Research in Rockville, Maryland, had taken the smallest known bacterial genome, that of *Mycoplasma genitalium*, and started to knock out its genes to find out which were essential to survival. They predicted that a set of about 300 genes would be sufficient to maintain a single-celled organism in the laboratory.

Arthur Caplan of the Center for Bioethics at the University of Pennsylvania, who led the panel that reviewed the project, believes the creation of artificial life is a "certainty". The panel's review¹⁵ acknowledged the bioweapons potential of the minimal genome project and other genomics efforts, concluding that "we need to give serious thought to monitoring at the level of national and international public policy".

Synthetic pathogens — probably viruses, rather than bacteria — are on the list of long-term threats being considered by experts in biodefence. "It's not out of the question that we'll be able to produce artificial viruses," says Mark Wheelis of the University of California, Davis, a member of the Federation of American Scientists' working group on bioweapons.

But even if 'artificial' pathogens can be created, most experts are sceptical of their relevance to biowarfare — at least given our current understanding of ecological genetics. If genetically engineered bacteria and viruses would struggle to survive if released into the environment (see main story), the likelihood of a completely synthetic pathogen prospering in the wild seems slim.

that boosts antibody production.

But the IL-4 gene also effectively shut down the cellular arm of the animals' immune systems, rendering them unable to fight off mousepox. Most disturbing was that mice previously vaccinated against the virus also succumbed, being killed within days¹. Given that mousepox is a relative of smallpox, the potential for using similar techniques to develop an enhanced bioweapon is all too obvious. "If one were to use a bioengineered virus of this nature, it may not be possible to vaccinate against it," says Ramshaw.

Such examples reveal that some developments in biology have nightmarish potential. But many experts say that at present the reality of the threat posed by bioengineered weapons is probably much less than that from conventional biological agents. "The worst that you can imagine is probably not a very realistic scenario," says Albert Osterhaus, a virologist at Rotterdam University in the Netherlands. One reason for optimism is that pathogens engineered in the lab may struggle to survive, or quickly lose their imbued characteristics, if they were ever released. Evolution, argues Ewald, is on our side. "People don't think about natural selection. If they did, they would have a clearer idea of what the dangers would be."

Trading places

Because evolution is all about trade-offs between the costs and the benefits of different traits in particular environments, Ewald suspects that it would be extremely difficult to engineer all of the desired 'attributes' into a bioweapon and still have an organism that is transmitted effectively and predictably. In naturally occurring pathogens, he points out, traits such as virulence and transmissibility often counteract one another¹³.

Microbial evolution is also usually a matter of use it or lose it. Traits such as toxin production usually impose costs on their bearers, and so are likely to be lost quickly by engineered organisms unless they confer a selective advantage. The key question is how much damage an engineered pathogen might inflict before losing its added genes.

Even if introduced genes can persist, pathogens grown in culture tend to adapt to their new environment and lose the characteristics that made them pathogenic. Indeed, this strategy has been used to create harmless strains of viruses, such as polio, for use in vaccines. A bacterium engineered in culture to resist antibiotics may soon become similarly benign, suggests Stemmer: "You may end up with something antibiotic-resistant but no longer pathogenic."

The sinister implication is that the most effective means of developing an enhanced bioweapon would be to use human subjects in its development and testing. "If anyone were to use humans as guinea pigs, they could make some very nasty weapons," says Ewald.

The worst that you can imagine is probably not a very realistic scenario.

Such horrific possibilities underscore the need to strengthen the 1972 international Biological Weapons Convention, say experts.

Counterstrike

Despite the evolutionary arguments against assuming a worst-case scenario, researchers working on biological defence take the threat of bioengineered weapons seriously, and are trying to develop the means to counter it. "Most detection technology is based on knowing what you are looking for," says Duane Lindner, who works on biodefence at Sandia National Laboratories in Livermore, California. "You could imagine engineering a pathogen so that detectors would be blind to that particular threat."

With this in mind, Sandia's Center for National Security and Arms Control is developing an Internet-based system to detect early signs of bioweapon exposure, irrespective of the agent responsible. The idea behind the Rapid Syndrome Validation Project is that doctors would enter details of unusual symptoms or disease outbreaks into the system. Neural network software would then be used to identify suspicious outbreaks of disease, without the need to wait for diagnostic laboratory data (see News, page 228).

In collaboration with the Lawrence Livermore National Laboratory in California,



Sandia is also trying to develop generic methods to detect biological agents without needing to know their identity. Lindner and his colleagues have used computational techniques to identify conserved regions of biological toxins, and they are now moving on to other proteins involved in pathogenicity. The knowledge gained could then be used to develop sensors to detect these molecular signatures.

The US Defense Advanced Research Projects Agency (DARPA), meanwhile, is developing biosensors based on living tissues that should provide physiological responses to a wide spectrum of both known and unknown pathogens. The biosensors are three-dimensional matrices containing cells including neurons, muscle cells, immune cells, and cells from the skin and the endothelia that line our guts and nasal passages.

DARPA is also investing heavily in the development of new antibiotics and vaccines that could target a broad range of pathogens. Some of DARPA's strategies target common mechanisms of bacterial growth, such as genes essential to cell division and those encoding enzymes central to evolutionarily conserved metabolic pathways. Maxygen, with funding from DARPA, is applying its DNA shuffling technology to combine proteins from related pathogens in the hope of developing vaccines that could provide broad protection.

In other words, the techniques that could produce bioweapons are also being deployed to set up countermeasures against them. This neatly illustrates the point that legitimate and malevolent applications of biology are merely two sides of the same coin.

Although bioengineered weapons may currently be less of a concern than their conventional counterparts, the threat they pose can only increase as technologies develop. "It's time for biologists to begin asking what means we have to keep the technology from being used in subverted ways," says Matthew Meselson, a molecular biologist at Harvard University who has spoken out frequently on the dangers of biowarfare. ■

Carina Dennis is a senior biology editor with Nature.

1. Jackson, R. J. *et al. J. Virol.* **75**, 1205–1210 (2001).
2. Block, S. M. in *The New Terror: Facing the Threat of Biological and Chemical Weapons* (eds Drell, S. D., Sofaer, A. D. & Wilson, G. D.) 39–75 (Hoover Institution Press, Stanford, 1999).
3. Stover, C. K. *et al. Nature* **406**, 959–964 (2000).
4. Cole, S. T. *et al. Nature* **393**, 537–544 (1998).
5. Heidelberg, J. F. *et al. Nature* **406**, 477–483 (2000).
6. Cole, S. T. *et al. Nature* **409**, 1007–1011 (2001).
7. Perna, N. T. *et al. Nature* **409**, 529–533 (2001).
8. Kuroda, M. *et al. Lancet* **357**, 1225–1240 (2001).
9. Stemmer, W. P. C. *Nature* **370**, 389–391 (1994).
10. Stemmer, W. P. C. *Proc. Natl Acad. Sci. USA* **91**, 10747–10751 (1994).
11. Cramer, C., Raillard, S.-A., Bermudez, E. & Stemmer, W. P. C. *Nature* **391**, 288–291 (1998).
12. Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. *Microbiol. Rev.* **56**, 152–179 (1992).
13. Ewald, P. W. *Evolution of Infectious Disease* (Oxford Univ. Press, New York, 1994).
14. Hutchinson, C. A. III *et al. Science* **286**, 2165–2169 (1999).
15. Cho, M. K., Magnus, D., Caplan, A. L., McGee, D. & the Ethics of Genomics Group *Science* **286**, 2087–2090 (1999).

A fine set of threads

Ultra-thin fibres spun from polymers could be used to protect against chemical weapons, dress wounds and make brakes for aircraft. David Adam tells a gripping yarn.

The best-dressed soldiers of the future may not be wearing standard camouflaged battle fatigues, but a daring little number that Heidi Schreuder-Gibson likes to call “a head-to-toe odour-eater”. By reviving a 70-year-old process called electrospinning, Schreuder-Gibson hopes to create a breathable fabric that can trap and deactivate lethal nerve gases.

Current chemical-weapons suits are relatively crude ensembles consisting of little more than Wellington boots and rubberized trousers and tunics. Such barrier protection keeps gases out, but also retains body heat — Gulf War troops found the suits stiflingly hot. Schreuder-Gibson and her colleagues at the Natick Soldier Center in Massachusetts may have the solution. Their fabrics are made from threads just a few nanometres wide that are ‘spun’ from drops of liquid polymer. Speaking to a packed session at the American Chemical Society’s meeting in San Diego last month, Schreuder-Gibson described how she mixed catalysts capable of breaking down nerve gases into the drops of polymer. Fabrics made from the resulting fibres blocked liquid versions of nerve agents such as mustard gas and sarin.

Electrospinning’s potential does not end there. Mix a substance that protects against infection into the polymer drop and the fibres could be used to create materials for wound dressings. Other researchers want to mix in compounds that release drugs over periods of time.

The technique dates from the 1930s. “The big textile companies had a look at it, decided it didn’t offer anything new and put it aside,” says Darrell Reneker, a polymer scientist at the University of Akron in Ohio. But the rise of nanotechnology persuaded Reneker and others to take another look. “We realized that what this technology offered was access to nanoscale diameters,” he says.



Mighty mesh: built from fibres that make a human hair look bulky (left), the fabric (above) withstood liquid versions of nerve gas in tests.

By applying tens of thousands of volts, it is possible to overcome the surface tension that holds liquid drops together. The voltage pulls charge from inside the drop onto the surface, where it causes surface molecules to repel each other. These push away from each other, allowing new molecules to come to the surface. This makes the drop very unstable. Controlling the process is tricky, but, under the right conditions, the drop can become a rapidly growing polymer jet that follows a tightly looping spiral path. The jet’s width shrinks to as little as three nanometres as its length increases.

Synthetic skins

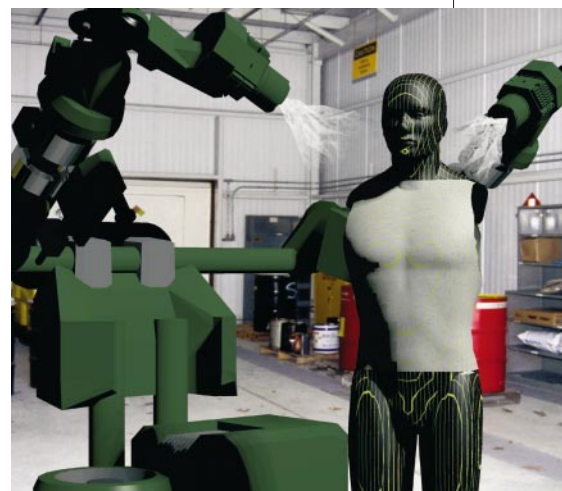
Mats of polymer nanofibres can then be built up by collecting the rapidly solidifying jets on a surface. The technique works with over 30 different organic polymers, as well as some non-polymeric materials including carbon and certain ceramics. Electrospun fibres can cost up to 30 times as much as nanofibres produced by drawing out strings of polymer from solidifying polymer solutions. But this has not stopped eSpin Technologies, the world’s only electrospinning company, based in Chattanooga, Tennessee, from shipping fabrics to researchers investigating their potential as synthetic skins, scaffolds for artificial organs and brake pads for aircraft. Customers are prepared to pay because electrospun fibres can be made one-tenth the thickness of conventional nanofibres. Researchers also like the ability to tweak the chemistry of the fibres and to con-

trol their diameter accurately, says eSpin founder Jayesh Doshi.

For Schreuder-Gibson, this difference in fibre width is critical. The thinner the fibre, the finer the mesh of fibres in the fabric, and the greater the chance of catching and breaking down an incoming molecule of nerve gas.

Ultimately, she hopes to create whole suits by electrospinning fibres onto mannequins, a process she estimates will take 10 years to develop. She also has one more trick up her sleeve. By blending carbon nanotubes and conductive polymers into the mix, Schreuder-Gibson’s team have made fabrics that conduct electricity. Nerve gas sensors woven into the suit could then detect the gas and warn the wearer — making it the first odour-eater to neutralize offending chemicals while complaining about the smell. ■

David Adam is a news and features writer for *Nature*.



Suits you: ten years from now, garments may be made by spraying fibres on to a mannequin.

Hero or villain? Stasi archives shed light on Russian scientist

Sir — Nikolai Timoféeff-Ressovsky (1900–81), one of the most striking personalities in twentieth-century science, did ground-breaking research in the fields of population genetics, radiation biology and evolutionary biology while working in Germany and the Soviet Union (see, for example, D. Paul and C. Krimbas, *Scientific American* **266**, 86–92; 1992). Until now, studies of his life, which began in Tsarist Russia, have been hindered by language barriers, the cold war and inaccessible archives. We have recently unearthed material for the first time from the archives of the Stasi (East German security service) which illuminates some, although not all, of the questions surrounding his life.

The material includes Third Reich material from Timoféeff-Ressovsky and his family; records of interrogations by Soviet officials after the Second World War of him and his German colleagues Karl Zimmer and Hans Born; an official 1988 Soviet investigation into whether he should be rehabilitated; and an East German investigation provoked by a biography.

Soviet officials had accused Timoféeff-Ressovsky of treason on three grounds: failure to return from Germany after going to do research there in 1925; providing Germany with information on Soviet scientific institutes; and contributing to the German war effort. Timoféeff-Ressovsky, who had been running an independent research institute in Berlin since 1937, denied working for the German war effort, although other scientists in his institute had done so.

Zimmer told the Soviet officials that wartime research at the institute had included the biological effect of neutron radiation; the manufacture of radioactive elements, including radium; the effect of X-rays on humans; paints to illuminate instruments in aircraft; X-ray weapons against enemy planes; the effect of cosmic radiation on pilots at high altitudes; and protection from radiation. Zimmer also testified that, beginning in 1939 for the Kaiser Wilhelm Institute for Brain Research, and in 1942 for Timoféeff-Ressovsky's genetics department, war work was carried out, including research on "weapons of mass destruction": X-rays and neutron radiation.

Born described radiation experiments on animals, on volunteers (including Born himself) and on human corpses. The transcripts of these interrogations make clear that the Soviet security service was mainly interested in military research and

experiments with radium and uranium. Afterwards, Timoféeff-Ressovsky spent nearly a year in a prison camp.

During the period of *glasnost*, Daniel Granin's book *Zubr* (Novyj Mir, Moscow, 1987) portrayed Timoféeff-Ressovsky as a scientific genius victimized by stalinism and as an anti-fascist who fought against Hitler: one of his sons had died in a Nazi concentration camp. Perhaps encouraged by this, Timoféeff-Ressovsky's surviving son Alexei sought his father's rehabilitation. Soviet justice officials rejected this request in 1988, on the grounds that his father was a traitor who had worked on weapons of mass destruction for Germany.

In 1989, by contrast, East German officials from the Stasi, the ruling Socialist Unity party and the Academy of Sciences noted that Timoféeff-Ressovsky had only given information to Germany on Soviet institutes during the German–Soviet non-aggression pact, when scientists had been encouraged to cooperate, and that he had stayed away from the Soviet Union to avoid persecution for opposing the then state-approved theory of lysenkoism. They concluded that the war work done at his institute came to nothing. This report by the East German Academy of Sciences may explain why Soviet officials rehabilitated Timoféeff-Ressovsky soon afterwards.

These sources suggest that Timoféeff-Ressovsky did not collaborate with the Third Reich for the war effort. But, as the East German report noted, neither could he be described as an anti-fascist.

Uwe Hossfeld*, Mark Walker†

*Ernst Haeckel House, Friedrich Schiller University, Berggasse 7, 07745 Jena, Germany

†Department of History, Union College, Schenectady, New York 12308-3163, USA

Sorting out the Smiths

Sir — 'J. Smith' published an astonishing 413 papers in the biosciences last year, with interests including radionuclides, retirement communities, HIV, endoglin (a TGF- β receptor-associated protein), childhood sexual abuse and its effect on the dating experiences of undergraduate women, rabies, t'ai chi and "Nice work, but is it science?"

Each of these papers can be accessed through the PubMed database by using the PubMed Unique Identifier (PMID) number — for example, 11232061 will locate the paper on dating experiences. But if you want to find, say, all papers published by the J. Smith working on HIV, you need to use additional qualifiers and/or know where the author has worked.

This suggests that the time may be ripe for the introduction of author-specific ID numbers, or AIDs, similar to PMIDs.

These could be provided at the time of publication and would form part of any reference database entry, making database searches considerably easier — though perhaps less eclectic.

Andrew Bradbury

Biosciences Division, MS-M888, TA-43, HRL-1, Building 1, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

Health-funding boost not enough for Canada

Sir — Your upbeat News (*Nature* **409**, 549; 2001) report of funds being poured into health research via the newly created Canadian Institutes for Health Research (CIHR) requires comment. Although it is true that the CIHR budget from April 1999 to March 2002 will be effectively twice that of the former Medical Research Council of Canada, Canada's expenditure on biomedical research remains paltry.

This year's CIHR budget of some Can\$477 million (US\$310 million) pales in comparison with the current budget of the US National Institutes of Health (NIH) at US\$20.3 billion. The population of Canada is one-tenth that of the United States, so the CIHR budget would have to be increased to about Can \$3.1 billion — US\$2 billion — to be comparable in per capita terms to that of the NIH. Because the NIH budget itself is on target to double over a five-year period, the United States will continue to outspend Canada by at least a 6:1 margin in the near future.

Although the creation of the CIHR and the doubling of biomedical research spending are very welcome developments, the truth is that operating-grant funding levels in Canada have improved from abysmal to simply inadequate. For example, only 54% of renewal grant applications and 25% of new grant applications were approved in the last CIHR competition, even though several decades of inadequate support had left only the best and brightest scientists competing for funds.

Moreover, the budgets of the 400 applications funded were reduced by an average of 12.8% from the minimum budgets recommended by the grant committees. The average size of a research grant was about Can\$92,000.

Until and unless the Canadian federal government makes several more doublings in the health-research budget, the statement by CIHR president Alan Bernstein — that Canada will be the place for health research in the twenty-first century — will continue to ring hollow.

Ronald N. McElhaney

Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2H7, Canada

Why science and religion need to talk

Human 'fallenness', freedom and pain are the preserve of both arenas.

SPL

Can a Darwinian Be a Christian? The Relationship Between Science and Religion

by Michael Ruse

Cambridge University Press: 2000. 272 pp.

£16.95, \$24.95

Paths from Science Towards God: The End of All Our Exploring

by Arthur Peacocke

Oneworld: 2001. 198 pp. £10.99, \$16.95 (pbk)

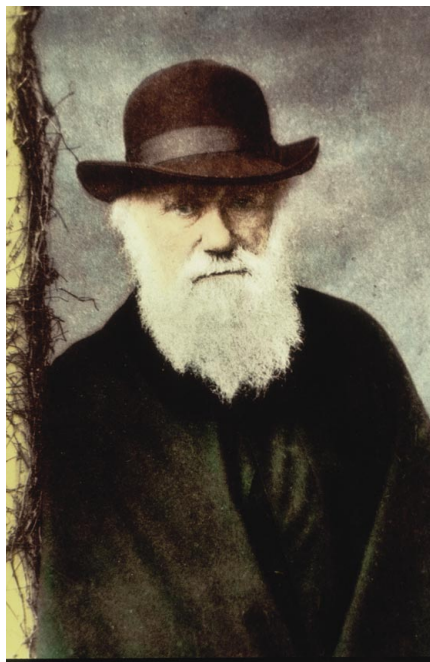
Charles L. Harper

Physics in the twenty-first century may indeed discover a 'theory of everything'. Particle physicists hope so. But many people may wish to reserve their unbridled enthusiasm. Discovering the unifying theory would be an astonishingly important human accomplishment; but it would not explain everything, even at a scientific level. Solid-state physicists and chemists might protest that the theory would not help them at all. Biochemists and biologists would ask where the origin and evolution of life fit in. A truly full-blooded 'theory of everything' would also need to address the 'Really Big Questions', which all intelligent people think about to some degree or other, and which traditionally have been the province of religion and philosophy. Do I have a purpose? Is death the end? Is there an 'ultimate reality'? Is there a God? What is love?

Science is wise to avoid these questions most of the time, yet it cannot ignore them completely. Increasingly, science touches on some of them in reasonably direct ways. The advance of darwinism especially, and its application to human evolution and psychology, has opened up the domain of engagement with the really big questions. This has also happened with discussions of the so-called 'anthropic principle' within cosmology.

Where these metaphysical discussions will go in the future remains an open question. Religious thought will play a natural and necessary adjunct role, retaining an important place at the table, whether as friend or foe, or simply as interlocutor. As science explores metaphysically rich terrain, it will be increasingly important for scientists to become conversant with the remarkably complex and interesting history of dialogue linking science and religion. *Can a Darwinian Be a Christian?*, by the distinguished philosopher of evolutionary biology Michael Ruse, is an illuminating and witty contribution to this literature.

A number of important books at the interface between darwinism and theology have been published in the past few years.



Authors of evolution: does Darwin's theory stress the natural evil of the world, opening a path to God?

These have mostly been aimed at people looking for a consonance between science and religion. Ruse's new book fills an important gap. It makes a case more appropriate to what might be called the core constituencies: the people with specific affinities — in this case with either darwinism or Christianity — rather than with 'science' in general or with 'religion' in general. These are people who are not necessarily uncomfortable with being non- or anti-Christian darwinians or non- or anti-darwinian Christians. They are not necessarily looking for consonance, and the challenge of dialogue offers no strong attraction. Why should such people care to read Ruse's book?

First, the issue of finding consonance is intrinsically interesting intellectually as well as spiritually. It engages the 'quests' of both science and religion for deep understanding of the world. Second, both science and religion have practical reasons for exploring constructive dialogue. Work in the history and philosophy of science has shown that common perceptions of a 'warfare relationship' between science and religion are largely myth-dominated. This situation tends to reinforce rather than ameliorate the conflict-perpetuating tendencies of both religious fundamentalism and its scientific counterpart, reductionist 'scientism'. Such tendencies are unhealthy for both science and religion.

For religion, continued conflict with science could have tragic consequences. As



the average standard of education rises and a higher proportion of the population participates in the knowledge economy, more and more people may dispense not only with fundamentalism but with religion altogether. As scientific knowledge grows, religious commitments predicated on 'gaps' in scientific understanding will invariably shrink as those gaps are closed. Those Christians who are currently fighting evolutionary science will eventually need to take it seriously; conflicts over philosophy are entirely another matter.

For science, the need to resolve the science-religion conflict has considerable strategic significance. In the United States especially, the persistence of conflict between Christianity and evolutionary biology impedes the broadening of public appreciation of science. This problem is particularly acute in view of the emerging genomics revolution, with its immense technological potential and the recurrent problems with the 'Frankenstein science' image. It also threatens support for the developing field of exoplanetary 'astrobiology'. Most broadly, religiously motivated antipathy towards science has a negative effect on the general climate for science education.

The third reason is that the future development of global civilization requires sensitive and thoughtful moral leadership on the control and utilization of the immense power emerging from science and technology. Both darwinians and Christians will be

BETTMANN/CORBIS

committed to such a future vision of responsibility — although often quite differently conceived. People's deepest cultural values, often nourished and embodied in religious contexts, should not be lightly cast into the dustbin of history as nothing more than outmoded nonsense. A certain humility and respectfulness by science towards the accumulated reservoir of moral wisdom contained in the world's religious traditions may be of considerable significance for the human future. Similar concessions by religious people towards appreciating the scientific heritage are also appropriate.

So Ruse's book serves an important role — building bridges for people who otherwise might not be interested in exploring 'win-win' as opposed to 'win-lose' relationships between science and religion. The argument is spry and engaging.

Ruse reasons fascinatingly that a Christian ought to find deep resonance with darwinism on major theological issues such as human "fallenness", freedom, pain and the problem of evil, and the importance of ethics. One section engages the issue of design, summarizing Ruse's ongoing debate over an anti-evolution movement, very popular with evangelical Christians in the United States, called 'Intelligent Design'. Ruse argues, quite rightly in my view, that the desire to ascribe the detailed particulars of nature (including anthrax, bubonic plague and cancer) to the explicit design of a Deity is a prescription for atheism, pointing precisely away from the Christian concept of God. Debating the conclusions of a much-cited anti-evolution book by the Intelligent Design movement biochemist Michael Behe, Ruse notes, "if Behe's argument actually points away from the Christian God, this should be acknowledged, for then Darwinism is surely a more attractive alternative for the Christian".

Ruse also takes to task several well-known popularizers of evolution who argue that atheism is an inevitable result of darwinism. "No sound argument has been mounted showing that Darwinism implies atheism. The atheism is being smuggled in, and then given an evolutionary gloss," he concludes. This topic covers some of the more significant passages in the book: these include a critique of the engagement of evolutionary history with St Augustine's attempted solution to the problem of evil, and a vivid endorsement of one of the twentieth century's most vital theological transformations from the classic view of the impassability of a transcendent being to the view of an immanent 'suffering God' bound into the torn fabric of the world's risky adventure. Ruse points out the sympathy of this view for the darwinian. "Darwinism stresses the natural evil of the world ... (and) opens the way for a Christian response", concluding that "Darwinism, a science which stresses physical suffering, looks to Christianity, a religion

which also stresses physical suffering and the divine urge to master it".

Readers interested in taking these themes forward will find much to stimulate their thinking in a new book by this year's winner of the Templeton Prize for Progress in Religion, Arthur Peacocke, a biochemist also trained as a theologian. *Paths from Science Towards God* ranges widely over relatively little-known topics where scientific ideas have been engaging academic theology in areas of considerable philosophical interest and sophistication. These include the concept of emergence, the mind-brain relation and top-down causality, evolutionary directionality and the evolution of altruism, the significance of spiritual experience, the challenge of religious pluralism, theistic naturalism, the conceptualization of divine presence and action in the world, and the significance of suffering and moral action.

Those looking for some of the best and most innovative thinking in 'science and religion' will find much of interest in this book, which is written at the level of a high-end introductory overview. Both Ruse and Peacocke cite with approval a judgement made in 1889 by the theologian Aubrey Moore: "Darwinism appeared, and, under the guise of a foe, did the work of a friend." After more than 100 years of rancour, it may be time to take Moore seriously. ■

Charles L. Harper is Executive Director of the John Templeton Foundation, Suite 100, Building 5, Radnor Corporate Center, 100 Matsonford Road, Radnor, Pennsylvania 19087, USA.

Hopping through the Mathiverse

Flutterland: Like Flatland Only More So

by Ian Stewart

Perseus Publishing: 2001. 320 pp. \$25, £17.99

Lisa Lehrer Dive and Andrew Irvine

Hollywood producers know that it's risky to make sequels. If the original hasn't broken any records, what's the point of revisiting old ground? And if it has, then living up to expectations is bound to be a challenge. In the case of *Flutterland*, however, Ian Stewart has produced a superb sequel to a century-old classic.

In 1884, a Victorian headmaster named Edwin A. Abbott released his now-famous tale of the inhabitants of Flatland. Although these two-dimensional creatures mostly lived their lives contentedly within the euclidean plane, they were at times both titillated and frightened by rumours of a third dimension.

Writing under the pseudonym A. Square,

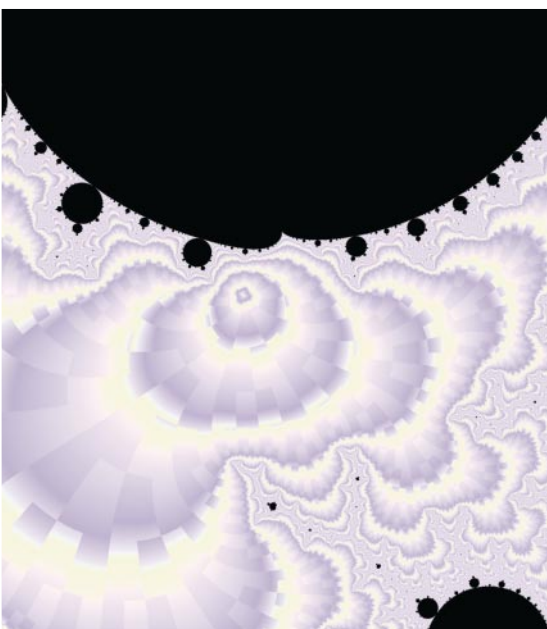
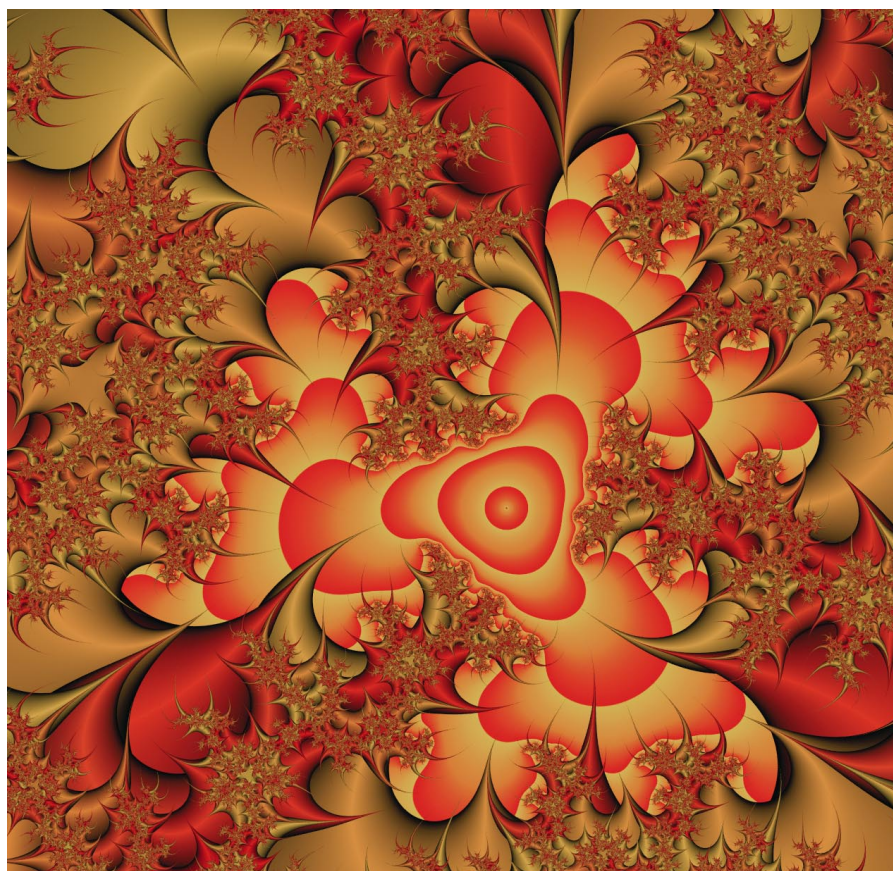
Abbott told the story of how Flatlanders learned to infer the shapes of friends and neighbours despite the fact that they were unable to visualize their world from above. (Just try distinguishing a circle from a square or triangle with your eye at table-top level; from the perspective of a Flatlander, apart from its width, every shape looks the same.)

Abbott also told of how a newcomer named 'Sphere' introduced the citizens of Flatland to the three-dimensional world of Space. When Sphere arrived, moving up and down through the plane, he appeared only as a circle growing and shrinking in size. From these scant data, Flatlanders had to infer the reality of the third dimension.

By analogy, the tale helped Abbott's readers understand the mathematics of a fourth dimension, something we three-dimensional creatures find equally difficult to conceptualize. At times both humorous and touching, the book was not only a good introduction to the mathematics of Abbott's day; it was also a thinly veiled commentary on the rigid social structure of nineteenth-century Britain and a warning of the dangers of original thought. The hero, who is the first to infer the existence of the third dimension, ends up in jail as a result of his heresy. Since its release more than a century ago, the book has never been out of print.

As its subtitle indicates, *Flutterland* by Ian Stewart is "like Flatland, only more so". In it, we follow the protagonist, Vikki Line, on her journey through space as she learns firsthand about the richness of contemporary geometry. Charming in both its simplicity and its humour, the book is briskly paced and gives non-mathematicians insight into some of the deepest and most fascinating mathematical structures developed over the past century.

Stewart adopts many of the same techniques used by Abbott in *Flatland*, and, like Alice in her journey through Wonderland, Vikki is transported to a variety of worlds in Stewart's (imaginary?) Mathiverse. With the aid of her guide, the Space Hopper, Vikki is introduced to a series of lively and appealing characters, including Mandelblot, the Taxi Controller of Quadratic City, and Moobius, the argumentative Cow. Together with Vikki, we visit several examples of prime surreal estate, including the Galois Fields, the Projective Plane and Topologica, the rubber-sheet continent. In Minkowski Space (which has been franchised out for the tourist trade) we meet the charmingly naughty space girls: Curvy Space, Bendy Space, Pushy Space, Square Space and Minny Space. And at the Topologist's Tea Party, Vikki chats with Schrödinger's cat Superpaws, a quantum animal that is subject to the Principle of Superpawsition. This leads to a discussion of several fascinating problems in which contemporary geometry and physics intersect.



Fractaland: a new way of taking a line for a walk.

just as the Flatlanders' conception of their world was inadequate for an understanding of what space is actually like, so too, our view of space as three 'flat' dimensions and a single temporal one is only the tip of the iceberg.

The narrative style of the book makes Stewart's ideas accessible and easy to digest, and Stewart himself makes ample use of analogies to help convey his underlying mathematical ideas. He also uses many examples to show just how fundamental mathematics is, and how everything around us, indeed everything we can think of, may be formed and discussed in mathematical terms.

Like Abbott, it is clear that Stewart knows the difference between serious science and solemn science. The book is as amusing and entertaining as it is informative, and serves as a fascinating introduction to the various mathematical spaces we inhabit.

Lisa Lehrer Dive is in the Department of Philosophy, University of Sydney, Sydney, New South Wales 2006, Australia.

Andrew Irvine is in the Department of Philosophy, University of British Columbia, 1866 Main Mall, Vancouver, British Columbia V6T 1Z1, Canada.

Just as Abbott used the Flatlanders' dream of Lineland and his subsequent visit to Spaceland to help his readers grasp the move from any given dimensionality to a higher one, Stewart uses his analogy with Flatland to show us new ways of expanding our own space. However, rather than merely introducing additional spatial dimensions, Stewart helps us understand the immensely complicated picture that modern physics gives of our world. Cleverly and engagingly, he opens our minds to the possibility that,

Affairs of the mind

Thinks...

by David Lodge

Secker & Warburg; 2001. 342 pp. £16.99

Steven Rose

David Lodge's campus novels have teased British academics for many years. Set in lightly fictionalized English universities, one or more powerful male academics and/or feisty females enjoy complex personal relationships and uncomplicated sex, providing the opportunity for shrewd and pointed observations about academic styles, and in particular, the pretensions of currently fashionable intellectual discourses. In *Thinks*, Lodge's new novel, the discourse under dissection is the latest enthusiasm among cognitive scientists — consciousness — and the setting is a greenfield university whose vast campus, with its separation of science and the humanities, serves as a metaphor for the 'two cultures'.

The alpha male, Ralph Messenger, is the charismatic director of an ultra-modern Centre for Cognitive Science, a telly-don who believes that consciousness can be incarnated — or better, inmachinated — in computers. His foil is novelist and lapsed Catholic Helen Reed, who is teaching a semester in creative writing. Messenger is engaged in a somewhat improbable introspective experiment, dictating his own stream of consciousness into his computer through a voice-recognition device (what he intends this to prove is never made clear). Reed keeps a traditional diary, and the two accounts work contrapuntally. Reed is taught to use e-mail, and the electronic courtship between her and Messenger is interspersed with the diary entries as they circle each other warily before tumbling into the inevitable copulatory frenzy.

The fun, for a neuroscientist at least, lies in Lodge's hilarious demystification of the pretensions of the new Artificial Intelligence, as Messenger attempts to convince a sceptical Reed of the legitimacy of computer consciousness. Told of Thomas Nagel's famous paper 'What is it like to be a bat?' and the parable of Mary the colour-blind colour scientist, Reed sets her students to write essays on the themes of being a bat, or Mary, in the deliciously parodied style of several well-known novelists.

The ironies multiply as Messenger's stream of consciousness meanders through academic politics, hopes of becoming a fellow of the Royal Society, sexual fantasies, and the entire subjective garbage heap that no computer is ever likely to share. This is James Joyce's Molly Bloom rather than Deep Blue or a Cray at thought.

Themes in the current debate are personified — from computationists through the new mysterians to quantum fantasists — through a description of scientists attending a conference on consciousness (the book should be required reading for Tucson Consciousness Conference addicts).

Lodge has done his homework; chief among the list of acknowledgements is his Birmingham colleague Aaron Sloman, although he avoids debating Sloman's own contention that computers can fall in love. As the novel insists, love and sex are human attributes. And consciousness is private — Messenger's affair with Reed collapses when he attempts to penetrate the subjectivity of her consciousness by reading her diary.

The novel is fun even if the plot is a bit clunky. But more than that, it is an elegantly and accessibly orchestrated introduction to the issues, fights — and, in light disguise, personalities — in current consciousness research. Just a couple of years ago, a Harvard neurophysiologist at a consciousness conference referred to the study of consciousness as a CLM — a career-limiting move. Not so any more, and Lodge has nicely tracked its arrival at academic respectability.

But Reed has the last word. At the final conference session, her lecture extols the joy of direct sensation with help from seventeenth-century metaphysical poet Andrew Marvell. We are left with the clear understanding that the ecstasies of sex, the pleasures of the taste of wine or of the ripeness of fruit are irreducible and non-computable. And if anyone is going to give us access to them, it is more likely to be a poet or a novelist than a cognitive scientist or intellectual artificer. Lodge has thrown down the gauntlet; it will be interesting to see who picks it up. ■

Steven Rose is in the Brain and Behaviour Research Group, The Open University, Milton Keynes MK7 6AA, UK.

Science in culture

Accessible architectures

Mapping the *Nature* website.

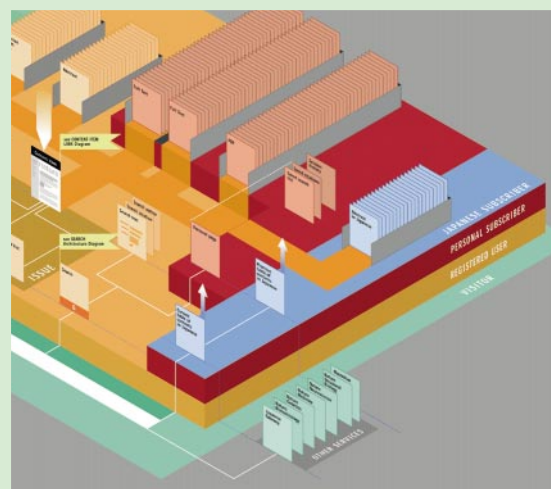
Martin Kemp

No major area of design can have undergone such a precipitate birth and manic growth as that of websites. In little more than ten years, a new design genre of enormous complexity — conceptually and graphically — has regaled us with an astonishing variety of visual solutions to the task of conveying digitized information. We have probably all experienced the serial frustrations of a poorly mapped site — “nominations on a postcard”, as they used to say in the pre-digital era — and most of us tend to take for granted the access provided by a well-conceived interface.

In either case, we probably give little thought to the conceptual and visual structures that lie invisibly behind the images and text on our screens. Indeed, we may suspect that many (perhaps most?) sites are designed without sufficient clarity at the vital stage when the architectural structure controlling visitor access should be visualized as a coherent whole. Paul Kahn and Krzysztof Lenk of Dynamic Diagrams, designers of the *Nature* website, nicely express the conceptual and perceptual parameters in their new book, *Mapping Websites* (RotoVision, £24.95, \$39.50):

“Building a website is not unlike building a physical structure, such as a house. One can start by planning the bedroom, then furnishing the kitchen and insisting on the best speakers for the stereo. But how do you get from the bedroom to the bathroom?”

They also point out that users with different interests and levels of expertise will require different kinds of plans: “The floor plan of Charlotte Airport on the back of the US Air Magazine contains enough detail for the passenger to visualize how to move from one gate to another. The floor plan of the same building required by a mechanical engineer will represent the same space with a different kind of detail.”

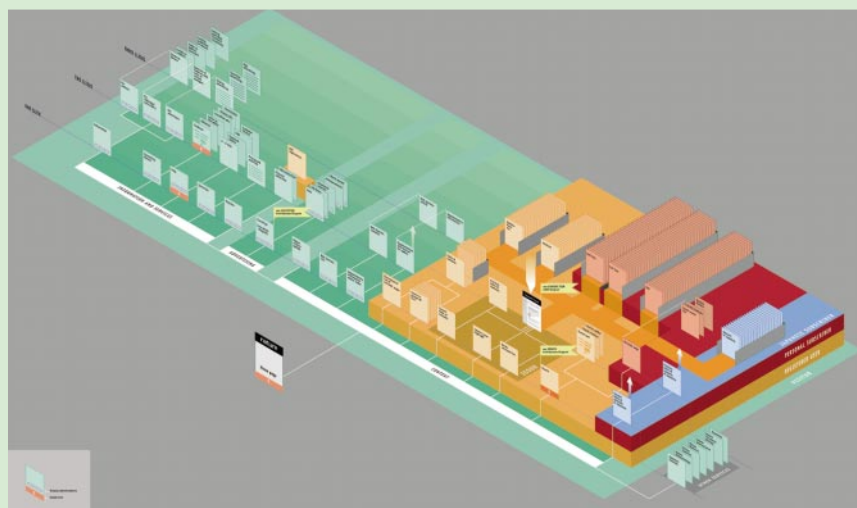


The process of a visualized three-dimensional space within which multidimensional information is systematically located is akin to the age-old techniques of the ‘art of memory’. Practitioners of the art, which reached its apogee in the sixteenth and seventeenth centuries, achieved astonishing feats of recall by secreting items in designated places in a familiar house or ‘palace’ of memory, so that they could look, say, under a chair or behind a particular door to find what they had previously hidden there. The technique relied on the astonishing power of association, long before modern studies of associational memory. In our era, the visual interfaces designed for our computers, with their icons, draw-down menus and other eye-catching graphics, all serve as readily recalled locators in our digitized palaces of memory.

The innovatory kind of planning map devised by Kahn and Lenk is exemplified by their evolved version of the overview of the *Nature* site. There are three dimensions: click depth (that is, the number of successive clicks needed to reach the appropriate depth in the structure); type of access (for example, by author, type of article or job advertisement); and, in the vertical dimension, hierarchy of access (visitors, registered users or personal subscribers, with a special blue floor for Japanese readers). Colour-coded cards and coloured ‘carpets’ contribute the further symbolic dimensions to the types and hierarchies of access.

The result of the architectural planning diagram, never seen by the visitor to the site, is to provide access to an intricately signed building through which we can progress with the minimum of fuss to see what is secreted in the room that interests us. Perhaps, on reflection, the underlying perceptual and conceptual principles are not so new, after all. ■

Martin Kemp is in the Department of the History of Art, University of Oxford, 59 George Street, Oxford OX1 2BE, UK.



Sense and sensibility

Could literature teach us how to release scientific writing from its straitjacket?

Robert Simmons

Lady Middleton could no longer endure such a conversation, and therefore exerted herself to ask Mr Palmer if there was any news in the paper. 'No, none at all,' he replied, and read on.

This passage from *Sense and Sensibility* by Jane Austen has several layers of meaning: it reveals something of the character of Mr Palmer; there is a gentle irony about the nature of newspapers and the reason for reading the newspaper; in its context it encapsulates upper-class England of 1800. It is a rich language that can imply so much in so few words.

For better or for worse, scientific language is not usually like that, particularly when it is confined by the straitjacket of the scientific paper. It was Peter Medawar who pointed out the fraudulent nature of the scientific paper, at least the standard experimental paper consisting of an introduction, methods, results and discussion. There may be scientists who logically consider past work in a field, then devise appropriate experimental tools, then do some experiments and then draw some conclusions, all in that order, but (*après* Austen) one would not wish to introduce them to one's daughters.

Medawar struck at the heart of the matter: in a scientific paper it appears that one section is derived from the preceding one by a process of induction. "The inductive format of the scientific paper should be abandoned," he wrote. "The discussion ... should surely come at the beginning. The scientific facts and scientific acts should follow the discussion, and scientists should not be ashamed to admit, as many of them apparently are ashamed to admit, that hypotheses appear in their minds along uncharted byways of thought; that they are imaginative and inspirational in character; that they are indeed adventures of the mind." Certainly, if this were common practice, the language of

science would be that much the richer.

The antithesis to this view is that writing a paper is a discipline to which we scientists all agree to subject ourselves, in which personal feelings are subjugated and the logic honed. We all (don't we?) try to bend the rules at times with covert professional fouls, pursuing personal vendettas disguised as honest criticism and so on, and it is the job of the editor to stop us. Of course a bit of personality is occasionally permitted to intrude — the more important the discovery and the more distinguished the scientist, the greater the indulgence. (There is Watson and Crick's calculated understatement "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible pairing mechanism for the genetic material"; and I remember a paper by Jeffries Wyman which started something like "While I was looking out of the window the other day, it occurred to me that...")

If the ultimate goal of science is objective knowledge, then should not the language be in some way 'superior' to the language of literature, which for the most part addresses the psychological and the subjective? Karl Popper described a framework for scientific knowledge and language in his 'three worlds': world 1, physical objects; world 2, mental experiences; world 3, products of the human mind (theories).

Human language, he wrote, "belongs to all three worlds. In so far as it consists of physical actions or physical symbols, it belongs to the first world. In so far as it expresses a subjective or psychological state ... it belongs to the second world. And in so far as language contains information [or] meaning ... it belongs to the third world. Theories, or propositions, or statements are the most important third-world linguistic entities."

Unfortunately (but perhaps wisely), Popper did not attempt to elaborate this scheme for the whole of knowledge or the whole of language. He did ascribe a higher place to the critical and descriptive aspects of language than to its communicative and expressive functions. In particular he seems to have argued that the critical aspect is a prerequisite for objective knowledge, which belongs to world 3. Whether his hierarchy really works for literature is an interesting question. All he seems to have said on the subject was that he set descriptive poetry (Homer, Shakespeare) on a higher level than lyric poetry (for example, Keats), which he held "is a sort of degenerative decline



No news is good news: do Mr and Mrs Palmer have something to teach scientific writers?

into completely expressive and communicative levels".

But whether the higher reaches of literature are lower in some hierarchy than, say, pure mathematics is arguable. Even in commonplace poems the text is often transcended: when John Masefield says "I must go down to the sea again, to the lonely sea and the sky, And all I ask is a tall ship and a star to steer her by" we know immediately that he is not after cod. When William Wordsworth's heart "dances with the daffodils" he is not talking horticulture. In Jane Austen the social setting is almost abstract: we watch a game with complex rules being played out.

There is also the question of originality. The arts and science alike thrive on novelty, but in most of science, however 'original' it is, we can only discover what already exists. Are we like Mr Palmer and the newspaper, just reading on, even though there is nothing new?

Robert Simmons is in the Randall Centre, New Hunt's House, King's College London, Guy's Campus, London SE1 1UL, UK.

FURTHER READING

Medawar, P. B. Is the Scientific paper a fraud? *Listener* 70, 12 September 1963. Reprinted in *The Threat and the Glory, Reflections on Science and Scientists* (ed. Pyke, D.) (Oxford Univ. Press, Oxford, 1990).

Popper, K. R. *Objective Knowledge, an Evolutionary Approach* (revised) (Oxford Univ. Press, Oxford, 1979).
Popper, K. R. *Knowledge and the Body-Mind Problem: In Defence of Interaction* (ed. Notturmo, M. A.) (Routledge, London, 1994).

Should not the language of science be in some way 'superior' to the language of literature?

Ecological forecasts

Gretchen C. Daily

Until the next big asteroid hits us, the future of life on Earth will depend much more on humanity than on anything else. Collectively, we have the power to wipe out most macroscopic species and to change radically the ecological and evolutionary playing fields of any that remain. Humanity exercises this power largely inadvertently, with consequences that increasingly threaten our well-being. What will happen from here on? Clues are being sought in the countryside, where in subtle but profound ways the possible futures of life are already unfolding.

'Countryside' refers to the growing fraction of the Earth's unbuild land surface whose ecosystem qualities are strongly influenced by humanity. It includes agricultural plots; gardens and pasture; plantation or managed forest; and remnants of native vegetation in areas devoted primarily to human activities. The countryside offers a broad arena where the functioning and fates of species can be studied under alternative predicted futures.

Yet surprisingly little is known about countryside biogeography, especially in rapidly changing tropical ecosystems. Initial studies and conservation efforts understandably focused on natural ecosystems and, more recently, on their remaining fragments. 'Island biogeography', a conceptual framework assuming islands of natural habitat surrounded by an inhospitable sea of agriculture and other development, informed much of the early work. Now, the 'sea' itself has become a focus of study as the fraction of the biosphere free from human influence is tiny

and rapidly shrinking; and many countryside habitats are actually not so inhospitable.

Three broad questions face us as wilderness fades away. First, assuming that human impacts intensify as currently projected, what sorts of species and ecosystems will exist over the coming decades and centuries? Second, what sorts do we want – and how do we decide? And third, how, and to what extent, can we achieve our desires? (See Box.)

Inspiring progress on all three questions has been made over the past two decades. New approaches and perspectives centre around building a science that will let us forecast changes both in biodiversity and in the way that ecosystems function. Many efforts aim to integrate ecology further with fields such as anthropology, climatology,

Seeking clues to the future

Scientific questions within the broad framework of biodiversity change include:

- Which species traits are advantageous in the face of major habitat alterations, and why?
- How will human-induced extinctions shape future diversity and evolution?
- What roles will countryside biotas play in supplying ecosystem services, such as water purification, flood control, pollination and preservation of options for the future?
- Are existing tools (ecological theory, remote sensing, interdisciplinary models for forecasting changes in land use) up to the task of predicting the persistence and functioning of countryside biotas?
- What practical measures can be taken to increase the capacity of countryside habitats to sustain biodiversity and ecosystem services?
- What are the trade-offs for society of alternative land-use regimes?

Countryside

"Humanity has always been, and always will be, a part of nature."

economics, history and law to characterize the consequences to society of 'ecosystem change' and possible responses to it.

There is great potential for biodiversity conservation, at least in the short term, in tropical countryside of 'low' to 'intermediate' intensity of use — small plots and paddocks (say, 0.1 to 10 hectares), and scattered remnants of native vegetation (0.1 to 100 or 1,000 hectares). For instance, more than half of Costa Rica's native bird species occur in largely deforested countryside habitats, together with similar fractions of mammals and butterflies. Little is known about long-term survival of biotas, however; work in Kenya and elsewhere shows that the path to extinction can be many decades or centuries long. Europe, much of which has been 'countryside' for a long time, shows clearly that some farming landscapes retain more biodiversity and valuable ecosystem services than others.

Much work is needed, and little time remains, to help steer a course that will at least protect the most vital of the Earth's life-support services. A three-pronged scientific effort is under way to serve this end. It has the daunting task of building models of ecosystem change, with emphasis on understanding the functional roles of species and ecosystems, especially in countryside. To preserve options, it must also develop more comprehensive strategies to conserve native biotas and their services on human-dominated lands, as well as in remaining 'natural' areas. Finally, there are interdisciplinary efforts to characterize the biophysical and socio-economic trade-offs associated with alternative regimes of land management — and to inform farmers, foresters and society generally about policy options. Nothing less than life is at stake. ■

Gretchen C. Daily is at the Center for Conservation Biology, Department of Biological Sciences, 371 Serra Mall, Stanford University, Stanford, California 94305-5020, USA.

FURTHER READING

- Daily, G. C. *Nature's Services* (Island Press, Washington DC, 1997).
- Daily, G. C., Ehrlich, P. R. & Sánchez-Azofeifa, G. A. *Ecol. Appl.* **11**, 1–13 (2001).
- Myers, N. & Knoll, A. (eds) *The Future Course of Evolution*. Special issue of *Proceedings of National Academy of Sciences* (in the press).
- Salzman, J. (ed.) *Protecting ecosystem services: science, economics, and policy*. *Stanford Environmental Law Journal* (in the press).
- The Millennium Ecosystem Assessment* <http://www.millenniumassessment.org>

In search of the future: countryside offers an opportunity to assess changes to ecosystems.

Leaf sensor for CO₂ in deep time

Wolfram M. Kürschner

How long have CO₂ and climate been linked? At least 300 million years, according to the results of a neat technique that exploits the relationship between concentrations of CO₂ in the atmosphere and pores in leaves.

As early as 1899, Thomas Chamberlin¹ proposed that changing concentrations of CO₂ in the atmosphere drove temperature and therefore climatic fluctuations during the most recent ice ages. A century after Chamberlin, analysis² of air trapped in ice cores provided direct evidence for a connection between atmospheric CO₂ and glacial–interglacial cycles that took the link back to 420,000 years ago. The link, of course, is that more CO₂ tends to mean higher temperatures.

As we go even further back into deep geological time, however, our knowledge of the CO₂–climate connection deteriorates rapidly. We mainly rely on modelling of the carbon cycle and on indirect evidence from various geochemical measurements for data that can inform us, by inference, of global mean temperatures and levels of CO₂. One group of authors³ has identified apparent inconsistencies in the CO₂–climate connection in an analysis that extends back 560 million years. They concluded either that CO₂ was not the main driver of climate for some one-third of this time, or that the reconstructions of CO₂ levels used were unreliable.

On page 287 of this issue⁴ Retallack describes how he has taken a different



Figure 1 A fossil *Ginkgo* leaf. Analysis of the stomata (pores) in leaves such as this allows reconstruction of the CO₂ environment in which the plant grew.

approach to the problem, using the stomata in leaves of fossil plants as biological sensors of past levels of atmospheric CO₂. Stomata are the pores on the epidermis of leaves that open and close to control gas exchange, including

uptake of the CO₂ used in photosynthesis. Retallack's approach is based on the inverse relationship, first noted by Woodward⁵, between the number of stomata on the leaves of woody plants and ambient concentrations of CO₂. He has created a new record of CO₂ concentration through time by gathering data from the leaves of living *Ginkgo* trees and from the fossil leaves of four groups of *Ginkgo* relatives (Fig. 1). The record is extraordinarily long, extending back to 300 million years ago.

Retallack validated his reconstruction of CO₂ levels in two ways. First, he used the measure 'stomatal index' (the percentage of stomata over stomata plus epidermal cells in leaves) rather than simple stomatal density. This minimizes possible confounding effects, such as different stomatal densities in plants of different sex, and the influence of environmental factors other than CO₂ — light, humidity and local temperature, for instance — that can likewise affect stomatal density. Second, he selected fossils only of those plant groups that show similar stomatal anatomy and indices, and that overlap in the fossil record.

Using experiments, and data from historical herbarium material, Retallack started by determining the relationship between stomatal frequency and CO₂ in modern *Ginkgo* leaves. He then applied this relationship to stomatal data from fossil leaf remains to

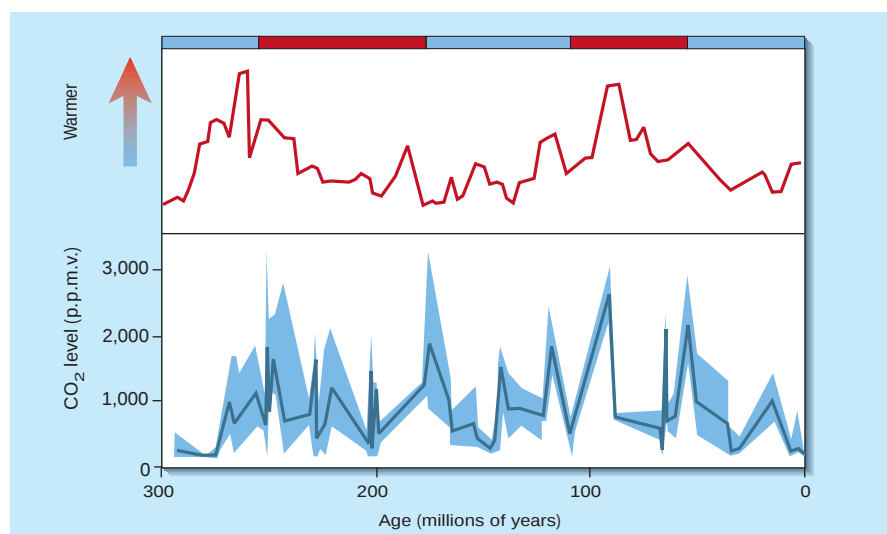


Figure 2 Levels of CO₂ in the atmosphere (bottom) and temperature trends over the past 300 million years. The CO₂ curve is the mean, with the 'envelope' showing the standard deviation, and is based on Retallack's stomatal-index analysis⁴ of *Ginkgo* leaves and the fossil leaves of closely related plants; p.p.m.v., parts per million by volume. Relative trends in temperature are inferred from the oxygen-isotope record of marine fossils³. Horizontal bars at the top indicate periods of predominantly warm (red) and cool (blue) modes of the Earth's climate, as inferred from the sedimentary record⁶ — glacial deposits in high-latitude regions, for instance.

reconstruct long-term trends in the changing levels of atmospheric CO₂ (Fig. 2). The general pattern ties in with the long-term temperature pattern established from oxygen-isotope ratios in marine fossils³. Periods of low CO₂ prevailed between about 296 million and 275 million years ago, between 30 million and 20 million years ago, and during the past 8 million years. The results agree with the predictions of carbon mass-balance models⁶ and, for the more recent part of geological history, with reconstructions inferred from marine geochemical data⁷. The periods of low CO₂ coincide with the periods of cool ('icehouse') modes of Earth's climate, as inferred from sedimentological evidence⁸ such as the extent of glacial deposits at high latitudes.

Warming trends are accompanied by increasing CO₂ levels between 275 million and 250 million years ago and from 130 million to 90 million years ago, whereas cooler periods from about 200 million to 140 million years ago and from 60 million years ago to the present are accompanied by reduced CO₂. The stomatal index record is incomplete, and its resolution is rather coarse (4 ± 6 million years). Nonetheless, compared with previous reconstructions of CO₂, Retallack's record shows many more maxima and minima that roughly correspond to the warmer and cooler periods seen in the marine temperature record.

In the older part of the record, before 150 million years ago, there seems to be a time lag of some 10 million years between shifts in CO₂ levels and in temperature. This lag may be a consequence of the poorer time resolution of the fossil-leaf record with increasing age, but it could of course reflect the influence of factors other than CO₂ on temperature. But all in all it seems to me that the new data point to a long-term coupling between CO₂ and temperature that is similar to the well-established coupling between the two during the most recent ice ages — roughly, the past 400,000 years. Moreover, the data seem to resolve the paradox of a cool climate and high levels of CO₂ at times during the Jurassic and Cretaceous periods (between 210 million and 120 million years ago). Retallack's results indicate that CO₂ levels were not as high as has been suggested by geochemical data⁹. A tendency¹⁰ to favour a decoupling of the CO₂–climate link may stem from flaws in models of the carbon budget and some CO₂ records, rather than incorrect temperature estimates.

It seems likely that this general technique — following trends in the stomatal index of closely related plant groups through time — might yield a valuable, better resolved record of CO₂ concentrations for a significant part of the Phanerozoic eon (that is, all the way back to 400 million years ago when plants first colonized the continents). But this approach does require further scrutiny. For example, at least for some parts of the geological record, do individual long-lasting plant groups

confirm the trends seen in a combination of different groups? Moreover, data with better temporal resolution are necessary to refine the present picture. The fossil-plant record is rather fragmentary compared with the marine temperature record; it will need much more work to see whether it contains enough clues to establish the exact temporal and causal relationship between CO₂ and climate on a timescale of several million years.

Concentrations of CO₂ will probably turn out to have been the main general determinant of temperature and climate during the past half-a-billion years of Earth's history. But it is likely that plenty of other factors came into play at different times and to different extents in the past. Changes in the configuration of continents, topography (as in mountain building) and ocean circulation,

or in Earth's orbit and the angle of its axis, as well as in solar brightness, can clearly all have a profound influence on climate. Teasing these various effects apart will be achieved only by interdisciplinary endeavour. ■

Wolfram M. Kürschner is in the Laboratory of Palaeobotany and Palynology, Utrecht University, Budapestlaan 4, 3584 CD Utrecht, The Netherlands. e-mail: w.m.kurschner@bio.uu.nl

1. Chamberlain, T. C. *J. Geol.* **7**, 545–584 (1899).
2. Petit, J. R. *et al.* *Nature* **399**, 429–436 (1999).
3. Veizer, J., Godderis, Y. & Francois, L. M. *Nature* **408**, 698–701 (2000).
4. Retallack, G. J. *Nature* **411**, 287–290 (2001).
5. Woodward, F. I. *Nature* **327**, 617–618 (1987).
6. Berner, R. A. *Am. J. Sci.* **294**, 56–91 (1994).
7. Pearson, P. N. & Palmer, M. R. *Nature* **406**, 695–699 (2000).
8. Frakes, L. A., Francis, J. E. & Syktus, J. I. *Climate Modes of the Phanerozoic* (Cambridge Univ. Press, 1992).
9. Ekart, D. P. *et al.* *Am. J. Sci.* **299**, 805–827 (1999).
10. Crowley, T. J. & Berner, R. A. *Science* **292**, 870–872 (2001).

Memory

Why is the cortex a slow learner?

John Lisman and Richard G. M. Morris

Mice that have only half the normal amount of a synaptic protein called α -CaMKII learn normally, but remember poorly. The result sheds light on the mysterious mechanisms of memory consolidation.

The gene-knockout business — engineering mutant animals with specific gene deletions — has turned up many oddities. Most are destined to be tangents on the path to discovery, but a skilful scientist can sometimes capitalize on unexpected findings. So it was with Frankland and colleagues' work on the involvement of an enzyme called α -calcium/calmodulin-dependent protein kinase type II (α -CaMKII) in neuronal plasticity and learning. As they describe on page 309 of this issue¹, they were studying α -CaMKII heterozygous mice, in which just one of the two copies of the gene encoding α -CaMKII was knocked out. Focusing on long-term potentiation (LTP), a molecular mechanism for strengthening neuronal connections that is widely held to be the basis of memory formation², the authors observed that this process was induced normally in both the hippocampus and the cortex in the brains of the heterozygous mice. But LTP then decayed unexpectedly quickly in the cortex, while persisting normally in the hippocampus.

These findings were brought to bear on the 'consolidation' theory of memory^{3,4}. According to one mechanistic explanation⁵ of this idea, newly acquired sensory information is funnelled through the cortex to the hippocampus. Surprisingly, only the hippocampus actually learns at this time — it is said to be online. Later, when the hippocampus is offline (probably during sleep), it replays stored information, transmitting it to the cortex. The cortex is considered to be a

slow learner, capable of lasting memory storage only as a result of this repeated replaying of information by the hippocampus. In some views, the hippocampus is only a temporary memory store — once memory traces become stabilized in the cortex, memories can be accessed even if the hippocampus is removed⁶. There is now direct evidence that some form of hippocampal replay occurs^{7,8}.

LTP is thought to be a molecular mechanism underlying learning, and requires the protein α -CaMKII (refs 9,10). This protein is greatly enriched at many excitatory neuronal connections (synapses) in the central nervous system. However, α -CaMKII heterozygous mice have only half the normal amount of α -CaMKII. Frankland *et al.*'s finding¹ that these mice show normal LTP in the hippocampus, but transient LTP in the cortex, allowed them to study consolidation in a new way. According to the theory above, the heterozygous mice should learn normally at first. But because LTP in the cortex is transient, information should not be stabilized there. The animals would be expected to forget quickly, and this is exactly what Frankland *et al.* observed. In two 'hippocampus-dependent' learning tasks, the mice learnt normally at first, but then forgot things more quickly than normal mice.

An elegant feature of the study¹ was the care taken, through further behavioural analyses, to explore alternative explanations of this apparent dissociation between initial learning and memory retention. One possibility was a 'ceiling' effect: the initial memory



100 YEARS AGO

In the *Revue Scientifique* of May 4, M. H. Coupin continues his essay on bird-song, dealing in this section chiefly with birds that imitate sounds other than their own. Very remarkable is the instance of a sparrow imitating the stridulation of the grasshopper. One spring a cage containing a sparrow was hung side by side with another in which were grasshoppers. No notice was taken by the sparrow of his neighbours, but next year, when he was again in the same society, he essayed the grasshoppers' chant. And for the rest of his life, when the grasshoppers had long been dead, the sparrow was accustomed to utter a polyglot song combining the notes of the insect with those of other birds. From *Nature* 16 May 1901.

50 YEARS AGO

Space-Time Structure. Schrödinger shares with Einstein the belief that there must be a natural generalization of the relativistic theory of gravitation which could serve as a secure base for dealing with all other natural phenomena: electro-magnetism, meson fields and their interactions. Einstein himself has given a short account of his ideas in a little book, "The Meaning of Relativity", which I have reviewed in *Nature* (166, 751; 1950). In fact, the two books by Einstein and Schrödinger have a common background of conviction and method; their aim is the same. But otherwise they differ thoroughly. Einstein's book discusses the idea of a unified theory only in a brief and condensed appendix, while the main text (a revised version of his Stafford Little Lectures of 1921, published 1922) is a rather elementary account of special and general relativity which can be recommended to students as an introduction to the subject. Schrödinger's new book is an attempt to prepare the reader right from the beginning for the generalized unitary theory. This is achieved by an extremely careful, and at the same time most elegant, development of tensor calculus in three stages... The mathematical foundations given in these chapters are a masterly work of art, a pleasure to read for a scholar or a well-trained student, but scarcely useful as an introduction for a beginner... This stimulating book can be recommended to all those who enjoy the intellectual pleasure of perfect logical penetration and condensed presentation of a theoretical structure derived from simple assumption, without demanding tangible results. Max Born From *Nature* 19 May 1951.

trace in the hippocampus might have been weaker in the heterozygous mice. However, when the authors trained the mice more intensively, the heterozygous animals still forgot more quickly. The implication is that trace strength and memory persistence are dissociable features of memory. This makes sense if they are, in part, stored in different parts of the brain.

These results support the idea that the hippocampus is the fast online learner that 'teaches' the slower cortex offline. But a close look at Fig. 3 on page 311 reveals an apparent contradiction: both the cortex and the hippocampus show rapid induction of LTP in response to brief stimulation, with LTP in the cortex remaining stable in normal mice. Why, then, should the cortex be a slow learner?

A possible solution relates to the fact that synaptic connections must already exist if they are to be strengthened by LTP. The most common form of LTP depends on NMDA-type receptors for the neurotransmitter glutamate to strengthen existing connections between neurons that fire simultaneously. As Hebb originally argued¹¹, such modifications could account for those types of learning in which associations are made between events or items that occur at the same time. In neuronal networks in the hippocampus, it is reasonably likely that there is a pre-existing synapse between any two cells, because the number of contacts a neuronal axon can make (10,000) is only slightly smaller than the number of cells in the network (generally less than a million)¹². So a sensory input would have a good chance of strengthening many synapses in the hippocampus by LTP. Networks in the cortex are much larger (within the 8-mm extent of horizontal cortical axons¹³, there are 10 million cells), and the probability of any two cells being connected is much lower. So there may be too few pre-existing connections that can be strengthened by LTP during learning.

But a connection between two cortical neurons might be formed and strengthened offline (during hippocampal replay) in a two-step process. The first step would be an ongoing process involving the perhaps random growth of new synapses and the breaking of older, non-stabilized ones. As the replay occurs repeatedly, when two cells do by chance become physically connected, they would then be sensitive to LTP if they were simultaneously excited as a result of hippocampal replay.

The observed development of new synapses in adult brains is consistent with this proposal. When animals are raised in mentally stimulating environments, the number of synapses increases¹⁴; studies of the mollusc *Aplysia* show a similar pattern¹⁵. More direct evidence comes from studies of horizontal axonal connections in the visual cortex of adult cats¹³ and in the monkey somatosensory cortex¹⁶. These studies show

that changes in the input activity to the cortex can lead to increases in axonal branching and in the number of synaptic boutons (terminals). This suggests that new connections can be made in adult brains, and that these connections are determined by processes that depend on neuronal activity.

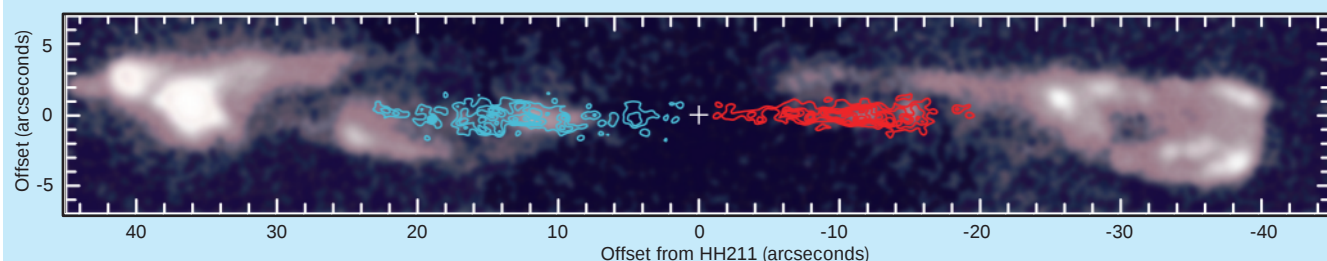
With the α -CaMKII heterozygous mice, it is now possible to manipulate cortical and hippocampal plasticity separately. It may also be feasible to settle some long-standing issues about how the processes in memory consolidation relate to retrograde amnesia. In humans, retrograde amnesia reflects the loss of access to, or the failure to consolidate, memory traces formed before brain damage. Animal models reveal a relationship between remembered events and the time before brain damage when those events occurred, providing hints to the underlying memory-consolidation process¹⁷. But such temporal gradients are not always seen, and certain features of retrograde amnesia can be understood instead in terms of the opportunities that recall allows for making several traces of one event, rather than the consolidation of a single trace¹⁸.

Timing the heterozygous deletion of α -CaMKII in mice would shed fresh light on this controversy. The mutant animals might remember information acquired shortly before the timed gene knockout. This would cast doubt on a slow, continuous interaction between brain areas and suggest that it is important that several memory traces are made in the hippocampus (in which LTP is normal in these mice). Alternatively, the mice might rapidly forget both old and new information. In either case, control of the mutation should enable us to probe deeper into the mysteries of memory. ■

John Lisman is in the Department of Biology and the Volen Center for Complex Systems, Brandeis University, Waltham, Massachusetts 02454, USA. e-mail: lisman@brandeis.edu

Richard G. M. Morris is in the Department of Neuroscience, University of Edinburgh, 1 George Square, Edinburgh EH8 9JZ, UK. e-mail: r.g.m.morris@ed.ac.uk

1. Frankland, P. W., O'Brien, C., Ohno, M., Kirkwood, A. & Silva, A. J. *Nature* **411**, 309–313 (2001).
2. Martin, S. J. et al. *Annu. Rev. Neurosci.* **23**, 649–711 (2000).
3. McGaugh, J. L. *Science* **153**, 1351–1358 (1966).
4. Squire, L. R. *Psychol. Rev.* **99**, 195–231 (1992).
5. Buzsáki, G. *Neuroscience* **31**, 551–570 (1989).
6. Eichenbaum, H. B. & Nadel, L. (eds) *Hippocampus* **11**, 1–60 (2001).
7. Skaggs, W. & McNaughton, B. *Science* **274**, 1870–1873 (1996).
8. Louie, K. & Wilson, M. *Neuron* **29**, 145–156 (2001).
9. Kennedy, M. B. *Nature* **335**, 770–772 (1988).
10. Lisman, J. *Trends Neurosci.* **17**, 406–412 (1994).
11. Hebb, D. O. *The Organization of Behaviour* (Wiley, New York, 1949).
12. Rolls, E. T. & Treves, A. *Neural Networks and Brain Function* (Oxford Univ. Press, 1998).
13. Darian-Smith, C. & Gilbert, C. *Nature* **368**, 737–740 (1994).
14. Klintsova, A. & Greenough, W. *Curr. Opin. Neurobiol.* **9**, 203–208 (1999).
15. Bailey, C. et al. *Nature Rev. Neurosci.* **1**, 11–20 (2000).
16. Flanagan, S. L. et al. *Science* **282**, 1062–1063 (1998).
17. Anagnostaras, S. G. et al. *J. Neurosci.* **19**, 1106–1114 (1999).
18. Nadel, L. & Moscovitch, M. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).



Astrophysics

A stellar performance

Kevin B. Marvel

Using a network of telescopes spread across the United States, astronomers have made a movie of an expanding shell of gas that sheds light on the intricate processes of how a star is born.

Our galaxy is a big place. Even the most fundamental objects in galaxies, the stars themselves, are of huge dimensions compared to human scales and are separated by even greater distances — the Sun has a diameter of 1.4 million km and its nearest neighbour, Alpha Centauri, is 41 trillion km away). Yet despite these vast distances, things change and move about on human timescales. As we orbit about the Sun, we move at an average speed of about 30 km s^{-1} , causing a slow but steady change in the positions of the stars we see each night.

On page 277 of this issue, an international team led by J. M. Torrelles¹ reports some striking observations of a shell of material expanding away from a young star in the Cepheus A star-forming region. The observations indicate that this material was ejected from the young star 33 years ago and is moving into the surrounding medium at the moderate speed of 9 km s^{-1} . The authors also observe that the material forms a spherical shell, contrary to models and observations of similar young stars, in which the ejected material forms two jets that flow in opposite directions.

A star is formed when a large, slowly rotating cloud of dust and gas collapses under the effects of gravity. Models of star formation predict that the jets of material formed near the poles of young stars carry away excess angular momentum (Fig. 1, above). The colour coding in Fig. 1 indicates the Doppler shift of the gas, with blue contours indicating motion towards the Earth, and red contours indicating motion away from the Earth.

A full understanding of the kinematics of an object requires not only observations of Doppler motion, but also measurements of proper motion — the motion of an object in the plane of the sky. Proper motions are exceedingly small because of the vast dis-

tances to celestial objects (the distance to Cepheus A is about 30.9 trillion km, for example). Astronomers divide up the sky into angular degrees, such that 90° is the distance from the horizon to a point directly overhead. Barnard's star has the largest measured proper motion, 10.3 seconds of arc per year (1 arcsecond is $1/3,600$ of a degree), whereas most stars move only a small fraction of an arcsecond during a year.

At the distance of Cepheus A, a velocity of 10 km s^{-1} completely in the plane of the sky leads to a proper motion of only 3 milli-arcseconds per year. Measuring such a small change of position requires extremely high resolution. The resolving power of telescopes increases directly as their diameter increases. But the size of telescopes is limited by cost and material strength, which can be overcome only by combining light from two or more widely separated telescopes — this is the basis of interferometry. One example of this design is the Very Long Baseline Array, a set of ten radio telescopes spread across the United States. It can produce images with resolutions as fine as 100 microarcseconds and can measure relative changes in position as small as 20 microarcseconds, or better, depending on observing frequency and source strength.

Not all objects can be observed using this technique, however. Only very compact, bright objects — or material seen against a bright background — are good targets for very-long-baseline interferometry (VLBI). Torrelles *et al.*¹ picked one of the best targets for their VLBI observations: astrophysical masers. Masers are volumes of molecular gas that can amplify radio waves because there are more molecules in an excited vibrational, rotational or vibrational-rotational state than in the ground state (this is known as population inversion). Water, methanol, SiO and OH can all produce bright and

Figure 1 Models of stellar formation indicate that jets of material form near the poles of young stars and move outwards, forming bipolar outflows, which carry away excess angular momentum. Observations of the young stellar object HH211 reveal two jets in the $2.2\text{-}\mu\text{m}$ molecular transition of molecular hydrogen (grey scale)¹¹. High-resolution observations of SiO, which can trace the boundary layers of the material, also show two jets moving in opposite directions (red contours indicate redshifted emission and blue contours blueshifted emission)¹². The small cross marks the likely location of the protostellar source (HH211) responsible for the formation of the jet. New interferometric observations¹ of a different source in the Cepheus A star-forming region suggest that this particular star is ejecting material in a spherical shell, rather than bipolar jets. If this source turns out to be a young star, this finding would bring current models of star formation into question. (Image courtesy of Claire Chandler.)

compact maser emission in astrophysical environments.

Astrophysical masers most commonly occur near both old and young stars. Observations similar to those of Torrelles *et al.* have shown that, in many young stars, water masers form and move in association with bipolar jets^{2–6}. The proper motions of masers are measured by making several observations of the same source over a period of time, which can be sequenced to form a movie. A sample movie showing the changing distribution of the SiO masers around a mature star has been produced⁷. Proper motions of masers around young stars, combined with their line-of-sight motions determined by Doppler shift, show that the masing material moves at speeds as high as $75\text{--}80 \text{ km s}^{-1}$ and often traces out shocked regions where the jet collides with the surrounding medium. In contrast, masers near old stars move more slowly ($10\text{--}30 \text{ km s}^{-1}$) and form spherical shells, although they rarely exhibit perfect symmetry^{8–10}.

The water masers observed by Torrelles *et al.*¹ seem to be associated with a circular distribution rather than a bipolar outflow,

indicating that they are part of an expanding spherical shell. Spread along arcs 0.4–70 AU in size (where 1 AU is the Sun–Earth distance), the masers are expanding at about 9 km s^{-1} and lie on a circle with a radius of about 60 AU. The deviation from a perfect circular fit is less than 0.1%. Because few masers lie away from the circular distribution, it is likely that the shell is quite thin. After determining the position of the centre of the circle, the authors consulted archival radio data and detected a very faint source, probably the young star powering the expanding shell. But the maser circle is not complete, perhaps indicating that other portions of the shell are disrupted by interactions with the surrounding medium or other nearby stars undergoing formation.

The uniform shape and the thinness of the maser arc in Cepheus A strongly suggest that the material originated in a single, short-lived ejection event. If the object associated with this maser distribution is a young star, as Torrelles *et al.* argue, this means that, at the very earliest stages of stellar formation, mass can be ejected in episodic, spherically symmetric and brief events. Further observations, looking for molecules associated with stellar jets such as molecular hydrogen or SiO, would help to prove that the source is young, but the necessary resolution is not yet available. The planned Atacama Large Millimeter Array, a large interferometer to be built through international collaboration in Chile, could help shed light on the exact nature of the source when it is completed.

Existing theories of stellar formation cannot explain this unique source, so there is a chance it could actually be a mature star. The current observations only reveal details about the masers themselves, not their stellar host, and so we cannot be entirely certain of the host star's age. Only time will tell if the source that has formed this symmetric shell of water masers will eject more material, perhaps allowing a sequel to the stunning set of images presented by Torrelles and colleagues.

Kevin B. Marvel is at the American Astronomical Society, 2000 Florida Avenue NW, Suite 400, Washington DC 20009, USA.
e-mail: marvel@aas.org

1. Torrelles, J. M. *et al.* *Nature* **411**, 277–280 (2001).
2. Gwinn, C. R. *Astrophys. J.* **429**, 241–252 (1994).
3. Leppanen, K., Liljestrom, T. & Diamond, P. *Astrophys. J.* **507**, 909–918 (1998).
4. Claussen, M. J., Marvel, K. B., Wootten, A. & Wilking, B. A. *Astrophys. J.* **507**, L79–L82 (1998).
5. Furiya, R. S. *et al.* *Astrophys. J.* **542**, L135–L138 (2000).
6. Patel, N. A. *et al.* *Astrophys. J.* **538**, 268–274 (2000).
7. <http://www.jb.man.ac.uk/~pdiamond/txcam44.gif>
8. Marvel, K. B. *Publ. Astron. Soc. Pacif.* **109**, 1286–1287 (1997).
9. Bowers, P. F., Claussen, M. J. & Johnston, K. J. *Astron. J.* **105**, 284–300 (1993).
10. Richards, A. M. S. Thesis, Univ. Manchester (1997).
11. McCaughrean, M. J., Rayner, J. T. & Zinnecker, H. *Astrophys. J.* **436**, L89 (1994).
12. Chandler, C. J. & Richer, J. S. *Astrophys. J.* (in the press).

Neurobiology

Snails, synapses and smokers

Dennis A. Dougherty and Henry A. Lester

The discovery of a protein that controls the transmission of nerve impulses in snails is significant in its own right. It also advances our understanding of the vertebrate neurotransmitter receptor that responds to nicotine.

On pages 261 and 269 of this issue^{1,2}, Sixma and colleagues describe how they identified a protein, found in snails, that controls communication between nerve cells; how they characterized that protein's properties; and how they analysed its structure on the atomic scale. This comprehensive set of studies is impressive in itself. Even more important, the structure reveals the essential features of a key region of the nicotinic acetylcholine receptor—the prototypical member of a group of proteins with diverse roles in vertebrate and invertebrate nervous systems. How can we be so sure that a small soluble protein from a snail is relevant to a receptor in the vertebrate central nervous system? First, the amino-acid sequences of the snail protein and of the relevant portion of the vertebrate receptor are similar. And second, the new structure rationalizes almost

every result from over 40 years of biochemical and electrophysiological studies of the vertebrate receptor.

Communication between neurons occurs at junctions known as synapses. When stimulated, the presynaptic neuron releases a neurotransmitter, such as acetylcholine at 'cholinergic' synapses. The neurotransmitter diffuses away, and some binds to receptors, which are large proteins in the membrane of the postsynaptic neuron. The nicotinic acetylcholine receptor not only has a binding site for the neurotransmitter, but also has a channel portion (Fig. 1a); when the receptor binds acetylcholine or other 'agonists' (including nicotine), the channel opens to allow ions to flow.

Sixma and colleagues¹ started by studying a cholinergic synapse in cultured snail neurons. They detected the new protein —

Micromachines

Strain against the machine

The world of micro-electromechanical systems (MEMS) is a tiny one, in which fully functional devices, ranging from electronic circuits to biosensors, operate on micrometre scales. Unfortunately, this world is still mostly a dream, because important elements such as hinges are not only difficult to make but also break easily. P. O. Vaccaro *et al.* address this issue by using elastic strain to make a hinge that moves itself into position (*Applied Physics Letters* **78**, 2852–2854; 2001).

The first generation of MEMS devices was limited to only two dimensions as they were formed by etching techniques. Making devices in three dimensions provides much more flexibility and functionality. This can be done by using hinges that rise above the base layer, known as the substrate. But previous attempts at making hinges involved many steps that

required manual or electrostatic intervention. Furthermore, these conventional hinges are prone to defects, tend to stick, and may wear out quickly as sliding parts rub against each other.

Vaccaro *et al.* have taken a simpler approach by bonding two different semiconductors (indium–gallium–arsenide and gallium–arsenide) together to form part of their MEMS device. These materials have different-sized unit cells in their crystal lattices, and the mismatch causes strain. When the semiconductor layers are released from the substrate by etching, they stand up by themselves (see picture). This effect is similar to the way the bimetallic strip works in thermostats — it flexes according to changes in temperature because of the strain between the two metals. The authors also deposited a reflective layer on top of the strain layer to



make the hinges into tiny mirrors.

These hinges appear to be fairly robust — they bend with ease and move back to their original position after bending. By creating hinges of different lengths, the authors can change the angle of the mirrors relative to the substrate. More generally, by varying the amount of indium, or by using different materials altogether, the authors can design mirrors with specified angles. And because there aren't any other moving parts, wear and tear shouldn't be a problem.

Josette Chen

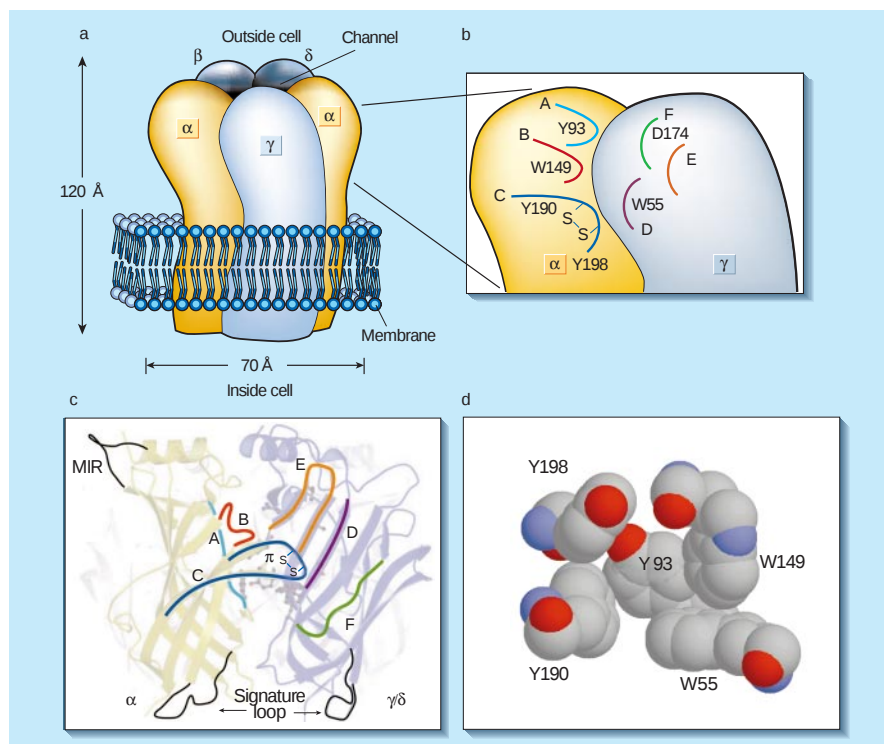


Figure 1 The nicotinic acetylcholine receptor. a, The overall layout of the receptor, based on cryoelectron microscopy studies by Unwin and colleagues⁵, showing the α , β , γ and δ subunits. This is the receptor found in postsynaptic muscle cells; some neuronal receptors consist of just one type of subunit. b, Biochemical results mapped onto the structure shown in a. The agonist-binding site consists of the interface between the α and γ subunits (or the other α and the δ subunit; not shown). One key disulphide bond is shown (labelled S-S), along with the six loops (A-F) that contribute to the agonist-binding site, and the key aromatic residues (Y, tyrosine; W, tryptophan). D174, aspartate residue 174. c, The results summarized in b are mapped onto Fig. 2b from ref. 2 (page 271). The agonist-binding site is denoted by π , emphasizing the importance of aromatic residues in general and tryptophan 149 in particular. Also highlighted are the signature loops of both subunits, and the main immunogenic region (MIR). d, The open box, formed by five aromatic residues, that probably defines much of the agonist-binding site. The view is derived from c by a 90° rotation clockwise around the vertical axis.

which they christen ‘acetylcholine-binding protein’ — by following the trail of an electrophysiological curiosity. They noticed that postsynaptic neurons at cholinergic synapses show much reduced responses to acetylcholine if glia (non-neuronal cells in the nervous system) are cultured nearby. This led the authors to purify the acetylcholine-binding protein, which they found to be released from vesicles in the glia. They sequenced this protein’s amino acids, cloned the gene encoding it, expressed the protein in yeast, and characterized both its effects in yeast and its distribution in snails.

This protein presumably decreases neurotransmission, but how? We know many quantitative details about the repertoire of molecules and mechanisms that make the nicotinic cholinergic synapse an exquisite electrochemical machine, specialized to function on a timescale of milliseconds and a distance of micrometres. For example, in order to open, the nicotinic acetylcholine receptor must bind at least two molecules of its agonist. Acetylcholine-binding protein might act as a buffer, decreasing the peak concentration of acetylcholine just after its release from the presynaptic neuron.

Acetylcholine would eventually be released from acetylcholine-binding protein, but asynchronously, so its concentration would decrease to the range where the likelihood of two agonist molecules finding the same receptor decreases dramatically. Some neurotransmitter transporters might have related buffering roles at various non-cholinergic synapses³. It is also possible that the glia release more acetylcholine-binding protein specifically in response to acetylcholine.

In effect, then, acetylcholine-binding protein would be a ‘decoy’ receptor for acetylcholine. The idea of decoy receptors is not new; indeed, one of the first artificial decoy receptors involved short peptides from the binding domain of the nicotinic acetylcholine receptor. These peptides sequestered nicotinic toxins, so protecting synapses⁴.

Sixma and colleagues² also describe the structure of the acetylcholine-binding protein, and this fits beautifully with biochemical and physiological observations of the agonist-binding region of the nicotinic acetylcholine receptor. Some of these earlier results are mapped onto a low-resolution image of the receptor⁵ in Fig. 1b. Pioneering work by Karlin and colleagues^{6,7} showed that there is an unusual structural feature in the α -subunit of the receptor, namely a disulphide bond between cysteines 192 and 193 (187 and 188 in the acetylcholine-binding protein). Another important advance was the identification by Changeux and colleagues⁸ of specific amino acids near the agonist-binding site.

These results, combined with those from mutation analyses^{9–11}, led to two general

Figure 2 Medical notes. The nicotinic acetylcholine receptor belongs to a superfamily that also contains receptors for the neurotransmitters serotonin, γ -aminobutyric acid (GABA), glycine and glutamate. The amino-terminal 230 or so amino acids of each receptor form the agonist-binding domain. This region, as well as binding neurotransmitters, also binds molecules of importance in human and veterinary medicine, as shown here. In addition, in myasthenia gravis (a human autoimmune disease), antibodies bind to the main immunogenic region of the nicotinic acetylcholine receptor. Finally, some naturally occurring mutations in the binding domain prolong the open state of the human nicotinic acetylcholine receptor in muscle. This leads to excessive activation, which damages the synapse and muscle fibre, causing a muscular weakness termed ‘slow-channel myasthenic syndrome’¹¹.

Receptor (number of known subunits)	Drugs and toxins that bind to the ligand-binding domain
Mammalian nicotinic acetylcholine receptor (≥ 16)	Nicotine, muscle relaxants, ganglionic blockers, possible painkillers, possible cognition enhancers, possible treatments for Parkinson’s disease
Mammalian serotonin 5-HT ₃ receptor (≥ 2)	Treatments for nausea, treatment for irritable bowel syndrome
Mammalian GABA _A -receptor family (≥ 19)	Benzodiazepines, muscimol (a toxin found in certain mushrooms), convulsants
Mammalian glycine receptor (≥ 5)	Strychnine (a convulsant), Zn ²⁺
Invertebrate glutamate chloride receptor (≥ 2)	Ivermectin (antiparasitic drug)

conclusions about the agonist-binding site. First, it is defined by several discontinuous regions of amino-acid sequence, known as loops. The main loops are A–C on the α -subunit and D on the γ - or δ -subunit, so the agonist-binding site spans an interface between subunits. Loops E and F are less important. Second, the agonist-binding site does not appear to contain a negatively charged amino acid to bind the positive acetylcholine, but instead is rich in aromatic residues (tyrosine and tryptophan). Apparently, acetylcholine binds the receptor through an interaction with aromatic residues¹², especially tryptophan 149 in the α -subunit (residue 143 in the acetylcholine-binding protein^{1,2})¹³.

When the earlier results are mapped onto the new structure² of the acetylcholine-binding protein, the remarkable image shown in Fig. 1c emerges. Loops A–D do indeed form a binding site, with loops E and especially F more remote. The disulphide bond is right in the middle of the action. The binding site is shaped by the five key aromatic residues, and resembles a box that is open at one end to allow the agonist to enter (Fig. 1d). The crystals of the acetylcholine-binding protein did not contain acetylcholine, but a molecule from the crystallization buffer was present, and an ammonium (positively charged) group from this molecule was positioned directly over tryptophan 143.

Another disulphide bond, between cysteine residues 123 and 136 (128 and 142 in the nicotinic acetylcholine receptor), produces a separate 'signature loop' that defines this group of proteins. But its position — the loop is at the very 'bottom' of the binding domain — is surprising. It means that, in the full receptor, the signature loop is positioned to interact directly with the membrane, or possibly with the transmembrane regions (or the short sequence connecting two of them) of the receptor. The implication is that the signature loop might be involved in 'gating' — the coupling of agonist binding to the opening of the ion-channel portion of the receptor. Intriguingly, the residues of the signature loop in the acetylcholine-binding protein interact more favourably with water than do those of the nicotinic receptors. This may be why this protein could be more easily overexpressed in soluble form for crystallization.

In the wake of these papers^{1,2}, computational models of the agonist-binding region of the acetylcholine receptor, as well as of those of other members of the group (the serotonin, γ -aminobutyric acid, glycine and glutamate receptors), will no doubt appear soon. Molecules that stimulate or block these receptors are useful in treating several ailments (Fig. 2), and structural information on the binding region will aid the design of even better treatments. Also, the structure of the transmembrane domain and, crucially,

the mechanism of gating must be worked out. And it remains to be seen whether the snail acetylcholine-binding protein has counterparts in mammals.

Dennis A. Dougherty is in the Division of Chemistry and Chemical Engineering, and Henry A. Lester is in the Division of Biology, California Institute of Technology, 1201 East California Boulevard, Pasadena, California 91125-2900, USA. e-mail: lester@caltech.edu

1. Smit, A. B. *et al.* *Nature* **411**, 261–268 (2001).
2. Brejc, K. *et al.* *Nature* **411**, 269–276 (2001).
3. Lester, H. A., Cao, Y. & Mager, S. *Neuron* **17**, 807–810 (1996).
4. Gershoni, J. M. & Aronheim, A. *Proc. Natl Acad. Sci. USA* **85**, 4087–4089 (1988).

5. Miyazawa, A., Fujiyoshi, Y., Stowell, M. & Unwin, N. *J. Mol. Biol.* **288**, 765–786 (1999).
6. Karlin, A. & Bartels, E. *Biochim. Biophys. Acta* **126**, 525–535 (1966).
7. Kao, P. N. *et al.* *J. Biol. Chem.* **259**, 11662–11665 (1984).
8. Corringer, P., Le Novère, N. & Changeux, J. *Annu. Rev. Pharmacol.* **40**, 431–458 (2000).
9. Pedersen, S. E. & Cohen, J. B. *Proc. Natl Acad. Sci. USA* **87**, 2785–2789 (1990).
10. O'Leary, M. E. & White, M. M. *J. Biol. Chem.* **267**, 8360–8365 (1992).
11. Sine, S. M., Quiram, P., Papanikolaou, F., Kreienkamp, H. J. & Taylor, P. J. *Biol. Chem.* **269**, 8808–8816 (1994).
12. Dougherty, D. A. & Stauffer, D. A. *Science* **250**, 1558–1560 (1990).
13. Zhong, W. *et al.* *Proc. Natl Acad. Sci. USA* **95**, 12088–12093 (1998).

Developmental biology

Making head or tail of Dickkopf

Roel Nusse

Signals that guide embryonic cells through development are often under the control of inhibitors. It now seems that one such inhibitor does not bind to the signal itself, but rather to the receptor that detects the signal.

A typical cell's network of signal-transduction pathways has so many molecular interactions that it looks like a complex wiring diagram. Recently, it has become clear that signalling events outside the cell can be equally elaborate, with many different components that bind to each other and act as positive or negative regulators of signalling. This is particularly true for proteins with key functions in development, such as bone morphogenetic protein¹, Hedgehog and Wnt. Various factors can interact with these proteins outside the cell, modulating their activity¹ or altering their

structure. On page 321 of this issue², Mao and colleagues describe an interesting twist to the regulation of extracellular signalling through Wnt proteins.

Wnt proteins, which are found in animals from hydra³ to insects, worms and vertebrates⁴, have a wide range of activities during animal development⁴; for example, they are involved in the formation of the head-to-tail axis of the embryo. Extracellular Wnt proteins trigger signalling pathways inside cells that proceed through several protein complexes that interact dynamically with each other.

One protein in these pathways is the

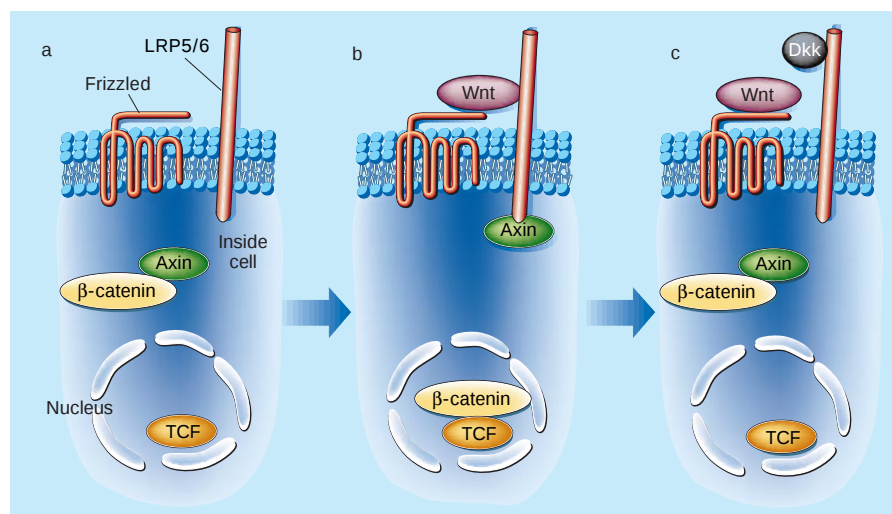


Figure 1 Events inside and outside the cell during Wnt signalling in development, based on recent results^{2,14,15}. **a**, In the absence of the Wnt protein, β -catenin is found outside the nucleus, in a complex with several proteins, including Axin. The transcription factor TCF is in the nucleus. **b**, The Wnt protein binds to its two receptors, Frizzled and LRP5/6 (for 'low-density-lipoprotein (LDL)-receptor-like protein 5 or 6'). Axin is then recruited to the intracellular tail of LRP5/6, releasing β -catenin in the process, which enters the nucleus to activate gene expression with TCF. **c**, Dickkopf (Dkk) binds to LRP5/6 and prevents the Wnt–Frizzled complex from interacting with LRP5/6. Axin is released and once more forms a complex with β -catenin.

β -catenin molecule⁵. Normally, β -catenin is kept in check by a large complex of several proteins (including one called Axin⁶; Fig. 1a). This complex promotes the addition of phosphate groups to β -catenin, enabling it to be detected by the cellular protein-degradation machinery. Signalling from Wnt releases β -catenin from its guards, allowing it to move to the nucleus, where it combines with a protein called TCF⁷ to activate the expression of target genes (including some that are involved in cancer⁵).

At the surface of cells, two kinds of protein are involved in receiving the Wnt signal. Members of the Frizzled receptor family consist of amino-acid chains that snake back and forth through the outer membrane of the cell. They use an extended amino-terminal region (called the cysteine-rich domain) to bind the Wnt protein⁸ (Fig. 1b). There are many genes encoding Frizzled proteins (ten in the human genome), and different Frizzled proteins probably have different affinities for various types of Wnt protein.

Genetic evidence from studies of fruit-flies (*Drosophila melanogaster*) and mice suggested that the second receptor for Wnt is LRP5/6 (refs 9,10). LRP5/6 is a long protein with a single transmembrane domain (Fig. 1b). Wnt proteins can form a complex with the cysteine-rich domain of Frizzled and with LRP5/6 (ref. 11), leading to a picture of a dual-receptor complex, which might form as the consequence of binding to Wnt. The intracellular parts of the receptors pass on this information, turning on the pathways that feed through β -catenin inside the cell.

It now appears² that LRP5/6 has a second function: it binds to a molecule that counteracts Wnt. That molecule is called Dickkopf (Dkk), meaning 'fat head', because it has the remarkable activity of promoting head formation in vertebrates¹². Active Wnt signalling can actually inhibit the development of anterior structures and the head¹³, and Dkk blocks this inhibitory effect in the appropriate parts of the embryo¹². The precise mechanism by which Dkk functions, however, remained unclear until now. Dkk and Wnt do not have similar amino-acid sequences, suggesting that Dkk does not act as a competitive inhibitor by binding to Frizzled or LRP5/6 in Wnt's place. In fact, it does not bind to Wnt or to Frizzled at all. Following Occam's razor reasoning, Mao *et al.*² tested whether Dkk binds to LRP5/6, and found that it does — through a part of LRP5/6 that is not needed for interactions with either Wnt or Frizzled. Similar results were obtained by Bafico *et al.*¹⁴. Binding to Dkk might alter the conformation of LRP5/6, so that it can no longer interact with Wnt and Frizzled (Fig. 1c). That would then halt the intracellular signalling pathways. But tests of this idea have been inconclusive² so far.

Another study has also placed LRP5/6 centre stage in Wnt signalling. Reporting

in *Molecular Cell*, Mao *et al.*¹⁵ show that the intracellular tail of LRP5/6 can bind to Axin (one of the proteins mentioned above that controls β -catenin in a Wnt-dependent manner). This observation provides a new link between the Wnt-receptor complex and intracellular components of these signalling pathways. Perhaps LRP5/6, once it has bound to Frizzled and Wnt, recruits and inactivates Axin, thereby releasing β -catenin (Fig. 1b). It will be interesting to see how the binding of Dkk to LRP5/6 influences the interaction between LRP5/6 and Axin. In general it is intriguing to find out whether the binding of different molecules to one receptor causes different events inside the cell.

These findings have several ramifications. For example, they may tell us something about the role of Frizzled in setting up the polarity of cells in a planar field¹⁶. Such planar polarity has been well studied in the outer layers of *Drosophila* tissues. Mutations in the *Drosophila* LRP5/6-encoding gene *arrow*⁹ or in the gene encoding Axin¹⁷ have no effects on planar polarity, so this process may not involve the binding of Wnt to the same dual-receptor complex. How can Frizzled have two separate tasks? The difference may lie in whether, once Wnt has bound to Frizzled, the complex then recruits LRP5/6 and Axin to signal to β -catenin rather than to set up polarity. Another question is whether Dkk is involved in Wnt signalling in *Drosophila*. There is no recognizable Dkk gene in the *Drosophila* genome¹⁸, but there could be unrelated proteins doing similar things. In fact, three other secreted factors inhibit Wnt in vertebrates, but none of these is detectable in the fly genome either¹⁸.

Ironically, the most mysterious player in this theatre is Wnt itself, as the protein has yet to be purified in an active form. But irrespective of these and other questions, it is clear that developmental signals are more complex than most secreted molecules that affect cellular function — probably because they have so many different jobs to do. ■

Roel Nusse is in the Department of Developmental Biology, Stanford University, and in the Howard Hughes Medical Institute, Beckman Center, Stanford, California 94305, USA.
e-mail: rnusse@cmgm.stanford.edu

1. Harland, R. M. *Nature* **410**, 423–424 (2001).
2. Mao, B. *et al.* *Nature* **411**, 321–325 (2001).
3. Hobmayer, B. *et al.* *Nature* **407**, 186–189 (2000).
4. Cadigan, K. M. & Nusse, R. *Genes Dev.* **11**, 3286–3305 (1997).
5. Bienz, M. & Clevers, H. *Cell* **103**, 311–320 (2000).
6. Zeng, L. *et al.* *Cell* **90**, 181–192 (1997).
7. Graham, T. A. *et al.* *Cell* **103**, 885–896 (2000).
8. Bhanot, P. *et al.* *Nature* **382**, 225–230 (1996).
9. Wehrli, M. *et al.* *Nature* **407**, 527–530 (2000).
10. Pinson, K. I. *et al.* *Nature* **407**, 535–538 (2000).
11. Tamai, K. *et al.* *Nature* **407**, 530–535 (2000).
12. Glinka, A. *et al.* *Nature* **391**, 357–362 (1998).
13. Kim, C. H. *et al.* *Nature* **407**, 913–916 (2000).
14. Bafico, A. *et al.* *Nature Cell Biol.* (in the press).
15. Mao, J. *et al.* *Mol. Cell* **7**, 801–809 (2001).
16. Adler, P. *Bioessays* **14**, 735–741 (1992).
17. Hamada, F. *et al.* *Science* **283**, 1739–1742 (1999).
18. <http://www.stanford.edu/~rnusse/wntwindow.html>

Daedalus

Watching the wind

The solar wind, says Daedalus, is a stream of electrons and positive particles emitted from the surface of the Sun at over 10⁵ metres per second. Modern electron optics is highly advanced these days, so it should be possible to obtain an image of the Sun's surface from its emitted electronic wind. So Daedalus is devising a satellite with a primary electron-coil a few hundred metres across, with a plane on which solar electrons can be focused. They will then emit light for photographic recording. Positive particles, repelled by the coil, should not interfere. In any case, Daedalus calculates that the Sun need only be a volt or so negative to emit electrons, whereas it must be about 2,000 volts positive to emit protons. He expects the solar wind to be biased in favour of the lighter particle. Daedalus also likes the idea of placing the satellite at the quasi-stable L1 lagrangian point, between the Earth and the Sun, where it can study the Sun continuously and relay its findings back to Earth.

The results, says Daedalus, should complement visual studies nicely. In normal times, the picture of the Sun by electron emission should match the visual one fairly well, although sunspots may appear anomalously dark (or bright), and emission near the poles may be dark as most of their electrons leave in other directions. The Sun is about 200 solar radii from the Earth. Assuming that the solar wind expands radially, the system has a fundamental magnification of about 200, which should enable fine detail of the Sun's surface to be resolved.

But the real value of the system will be in detecting magnetic storms and other unusual conditions. High-velocity streams in the wind may need special servo control of the electron-optic coil voltage to bring them to a proper focus; corresponding low velocities will need an opposite correction.

A portrait of a magnetic storm will be most intriguing. Will it be local or global? Good predictions should be possible. It may even turn out that fine resolution of the solar surface holds the key. But Daedalus's real goal is the electronic imaging of 'active' planets, such as Jupiter and Saturn, and even of detecting stellar winds from close stars. His satellite would have to be boosted to high velocity, of course, and its signal would be hard to detect. Daedalus reckons that stellar emission is much more intense at the poles, and some signal might even be received from the nearer stars this way.

David Jones

Sex differences in vocal learning in birds

Young females pick up songs much faster than males but are not so versatile later on.

A young songbird develops its songs by imitating adults¹, just as human infants acquire speech, listening to and memorizing adult songs during an early sensitive phase and then practising its vocalizations until they match the memory formed earlier². Here I investigate whether patterns of song learning differ between the sexes among northern cardinals (*Cardinalis cardinalis*), one of a few temperate species in which both sexes sing, rather than just the males. I find that females learn the same number of songs as males but in less than one-third of the time; however, auditory experience is not essential for males to develop their songs, whereas it is necessary for females to do so.

To determine the sensitive learning phase of male and female cardinals, nestlings (15 females, 11 males) were raised individually in an acoustically controlled environment and tutored daily by using tape-recorded cardinal songs over a period of 1 year. Each pupil was tutored daily with four tutor-song types which were changed regularly; in total, all pupils heard between 40 and 44 tutor-song types. At 1 year of age, a pupil's song was shown to be an imitation of a tutor-song type by their close spectrographic resemblance (Fig. 1a). Because tutor songs were played to the birds in a scheduled sequence throughout the year, identification of the copied tutor model allowed me to pinpoint when the bird memorized each song. The sensitive phase for each bird was defined as the entire period during which the bird heard tutor-song types that it later copied.

The sensitive learning phase was significantly shorter in females than in males: it started at a similar age in both sexes (males, 27.8 ± 2.80 days; females, 22.4 ± 2.06 days; Mann–Whitney test, $U = 43.5$, $P = 0.12$), but lasted more than three times as long in males as in females (males, 187.1 ± 33.4 days; females, 48.7 ± 9.8 days; $U = 11.0$, $P < 0.001$) and thus ended much later (males, 214.9 ± 35.3 days; females, 71.1 ± 10.5 days; $U = 17.0$, $P < 0.01$; Fig. 1b). The weighted mean age of song acquisition³, a measure that takes the number as well as the timing of songs acquired into account, was significantly younger in females than in males, with an average difference of 51 days (females, 40.1 ± 4.0 days; males, 91.5 ± 13.3 days; $U = 16.0$, $P < 0.01$; Fig. 1b).

The sex difference in the length of the sensitive phase cannot be attributed to exposure to an insufficient variety of tutor models, a factor known to delay the closure of the sensitive phase in other species⁴,

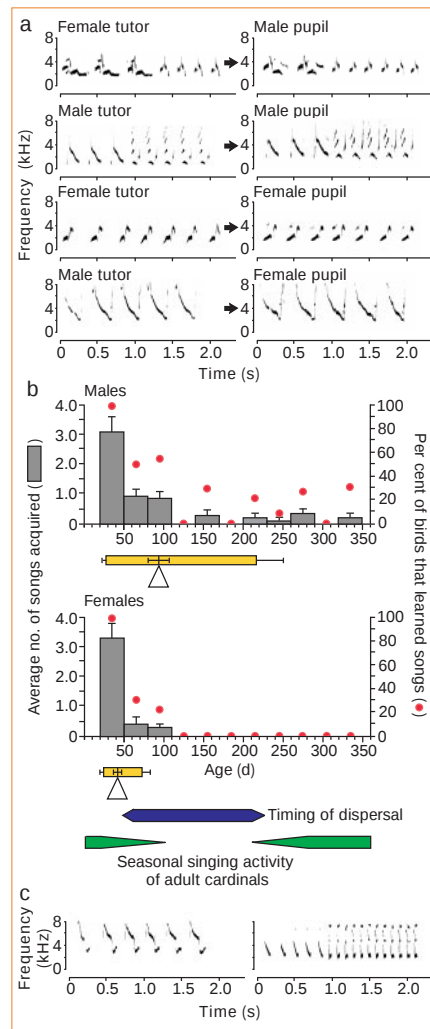


Figure 1 Gender differences in vocal learning in female songbirds. **a**, Sound spectrograms of tutor and pupil songs. Left, songs of male and female tutors: four song types with distinct acoustic profiles are shown. Right, attempted imitation of tutoring songs by male and female pupils. All the song components included in the pupils' songs closely resemble those of the tutor song types and are repeated in the same order as in the tutor songs. Male and female pupils imitated male and female tutor songs equally well. **b**, Timing of the sensitive phase for song memorization in male and female cardinals (15 females, 11 males). Bar charts show the mean and standard error of the number of songs acquired during each 30-day time bin by males (top) and females (bottom). Red points indicate the percentage of males and females that acquired songs during each 30-day period. Yellow bars indicate the average range of the sensitive phase (with standard error); triangles indicate weighted mean age of song acquisition. Blue bar, timing of dispersal in wild juvenile cardinals; green bars, seasonal singing activity of adult cardinals, indicating the availability of tutor models in the wild. **c**, Sound spectrograms of songs of two acoustically isolated males. Improvised songs developed by isolated male birds are composed of repeated song components that are characteristic of normal cardinal songs.

because males and females were exposed to the same variety of tutor models. Furthermore, when a subset of pupils was exposed to a richer variety of song types early in life (20 tutor-song types by 70 days of age for 2 males and 9 females), I found that the timing of their sensitive phases was still sex-specific and no different from that of pupils exposed to a regular variety of tutor songs (8 tutor-song types by 70 days of age for 8 males and 5 females; end of sensitive phase: U , 8.0 and 14.0, respectively; P , 0.99 and 0.26, respectively; weighted mean age: U , 5.0 and 10.0, respectively; P , 0.433 and 0.095, respectively).

Despite the difference in the length of the sensitive phase, the number of song types imitated was similar in both sexes (male, 5.2 ± 0.76 songs; female, 3.7 ± 0.45 songs; $U = 40.5$, $P = 0.08$). Although males acquired, on average, 1.5 more songs than females, the difference does not account for the threefold increase in the length of the

sensitive phase in males. In addition, although songs of male and female cardinal adults have subtle and consistent differences in acoustic morphology⁵, pupils did not show any preference for learning either male or female tutor songs: when exposed to both types (25% female songs, 75% male songs) during the sensitive phase, male and female pupils imitated songs of males and females equally well (sign test with predicted chance expectancy of 25%: females, $P = 0.227$; males, $P = 0.508$).

I also raised cardinals (2 females, 3 males) in acoustic isolation in order to determine whether a lack of tutor models could have sex-specific effects on song development. I found that isolated males developed improvised song types (5.0 ± 1.0 types) similar to normal cardinal songs (Fig. 1c). In contrast, isolated females sang only rarely, and when they did, their songs were of poor acoustic quality, indicating that auditory exposure to adult songs early

in life may be necessary to produce songs in females, but not in males. In summary, male and female cardinals differed markedly in when they acquired songs, but were similar in what they acquired. The sensitive phase estimated by using tape tutoring generally agrees with that measured using live tutoring⁶ and is likely to guide song acquisition in wild birds; the sensitive phase of wild male cardinals is entirely consistent with my results⁷.

The different sensitive phases for the sexes may be an adaptation to meet the functional requirements of songs. Juvenile cardinals disperse from their parents' territory to establish their own territory between the time they become independent from their parents (about 45 days of age) and the start of the first breeding season (when they are about 8.5 to 11 months old; Fig. 1b)⁸. My results show that, although both sexes start learning songs in their natal territory, the females lose the ability to acquire new songs soon after independence and males continue to learn songs until after dispersal is complete.

Songs of adult cardinals form 'dialects' in different geographical regions, so dispersing juveniles may settle into a population with a foreign dialect. The male-typical sensitive phase therefore allows him to

match his song types with those of his neighbours, even when he settles into a new population, whereas the female-typical sensitive phase preserves her natal dialect. Song matching may be important for male cardinals in the context of establishing territories, as in other species⁹, but evidently is not for females, for reasons yet to be identified. These findings parallel subtle but consistent gender differences in human speech acquisition¹⁰ and offer an opportunity to study the biological basis of distinctions in male and female learning.

Ayako Yamaguchi

Animal Communication Laboratory, University of California, Davis, California 95616, USA

e-mail: ay64@columbia.edu

Present address: Department of Pharmacology, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06520, USA

1. Marler, P. & Tamura, M. *Science* **146**, 1483–1486 (1964).
2. Marler, P. *J. Comp. Physiol. Psychol. Monogr.* **71**, 1–25 (1970).
3. Nelson, D. A., Marler, P. & Palleroni, A. P. *Anim. Behav.* **50**, 83–97 (1995).
4. Eales, L. A. *Anim. Behav.* **35**, 1336–1365 (1987).
5. Yamaguchi, A. *Condor* **100**, 504–511 (1998).
6. Nelson, D. A. *Anim. Behav.* **56**, 487–491 (1998).
7. Lemon, R. E. & Scott, D. M. *Can. J. Zool.* **44**, 191–197 (1965).
8. Smith, P. C. *Survival and Dispersal of Juvenile Cardinals*. Thesis, Univ. Western Ontario (1969).
9. Beecher, M. D., Campbell, S. E. & Stoddard, P. K. *Proc. Natl Acad. Sci. USA* **91**, 1450–1454 (1994).
10. Fenson, L. et al. *Monogr. Soc. Res. Child Dev.* **59**, 1–185 (1994).

Biogeochemistry

Phosphorus solubilization in rewetted soils

Biogeochemical cycles are shaped by events that follow soil drying and rewetting. Here we show that the process of drying and rapidly rewetting soil increases the amount of water-soluble phosphorus present and that this is predominantly in organic form after having been released from the soil microbial biomass. This effect could not only significantly affect phosphorus pollution of waterbodies but might also corrupt results from analyses involving water extraction of dried soils.

We investigated the effect of drying and rewetting soil on its content of soluble phosphorus by determining the amount of water-soluble phosphorus in 29 lowland permanent grassland soils from England and Wales. The total amounts of water-soluble phosphorus in moist soils were small, but increased after drying by 185–1,900%. Comparable amounts of phosphorus were solubilized when soils were dried at different temperatures, but solubilization occurred more rapidly at 30 °C (maximum phosphorus solubilization after 1–3 days) than at 15 °C (maximum solubilization after 7–10 days).

Most of the soluble phosphorus was pre-

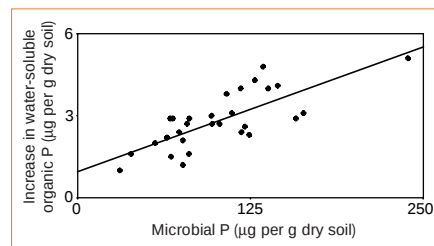


Figure 1 Increase in water-soluble organic phosphorus after soil drying as a function of soil microbial phosphorus. Data are from 29 permanent grassland soils in England and Wales containing various amounts of phosphorus (NaHCO_3 -extractable inorganic phosphorus was 9–48 mg kg^{-1}), total carbon (29–80 g kg^{-1}) and clay (22–68%). Increase in water-soluble organic phosphorus = $(0.95 \pm 0.33) + (0.018 \pm 0.003)$ (microbial phosphorus); $R^2 = 0.58$; $n = 29$; $P = 1.6 \times 10^{-6}$. The amount of water-soluble phosphorus was determined by extracting soils at field-moisture capacity (after drainage from saturation for 48 h) with water in a 4:1 water/soil ratio for 1 h. Subsamples were air-dried for 7 days at 30 °C and extracted in the same way. Extracts were filtered (pore size < 0.45 μm) and analysed for inorganic and organic phosphorus². Phosphorus levels in soil microbial biomass were determined by measuring the phosphorus released by chloroform fumigation⁶.

sent in an organic form, which accounted for 56–100% of the total phosphorus released by drying at 30 °C. The amount of solubilized organic phosphorus was positively related to the soil microbial phosphorus concentration ($P < 0.0001$; Fig. 1). We attributed this effect to direct release of phosphorus from the soil microbial bio-

mass, because microbes can be killed by osmotic shock and cell rupture when rapid rehydration follows a period of drying¹.

Phosphorus released from microbes after drying and rewetting the soil could contribute to the pollution of water bodies, because soil drying coincides with cracking, which enhances preferential movement of water and associated phosphorus through the soil. This may explain the increased concentrations of phosphorus found in drainage waters after periods of soil drying and rewetting². The impact of this effect on water quality may be increased by irrigation practices and in regions where climate change causes longer dry periods or more frequent cycles of wetting and drying³.

Our results also have implications for the usefulness of environmental soil-phosphorus tests that measure only inorganic phosphorus, rely on aqueous extraction, and take no account of soil moisture⁴. The effect shown here may also apply to other soil nutrients and could partially explain a similar phenomenon observed for nitrogen mineralization in tropical soils⁵.

Benjamin L. Turner*†, Philip M. Haygarth*

**Soil Science Group, Institute of Grassland and Environmental Research, North Wyke, Okehampton, Devon EX20 2SB, UK*

e-mail: phil.haygarth@bbsrc.ac.uk

†Present address: USDA–ARS, Northwest Irrigation and Soils Research Laboratory, Kimberly, Idaho 83341, USA

1. Salema, M. P., Parker, C. A., Kidby, D. K., Chatel, D. L. & Armitage, T. M. *Soil Biol. Biochem.* **14**, 15–22 (1982).
2. Turner, B. L. & Haygarth, P. M. *Soil Sci. Soc. Am. J.* **64**, 1090–1097 (2000).
3. Houghton, J. T., Jenkins, G. J. & Ephraums, J. J. *Climate Change. The IPCC Scientific Assessment* (Cambridge Univ. Press, Cambridge, 1992).
4. Pote, D. H. et al. *J. Environ. Qual.* **28**, 170–175 (1999).
5. Birch, H. F. *Plant Soil* **12**, 81–96 (1960).
6. Brookes, P. C., Powlson, D. S. & Jenkinson, D. S. *Soil Biol. Biochem.* **14**, 319–329 (1982).

Epidemiology

Foot-and-mouth disease under control in the UK

Following the first reported case on 20 February this year, foot-and-mouth disease spread to over 1,500 livestock farms in the United Kingdom by the end of April^{1–3}. From late March, the Ministry of Agriculture, Fisheries and Food (MAFF) required livestock on infected farms to be culled within 24 hours of the disease being reported and those on neighbouring farms within 48 hours. Here we investigate whether progress towards meeting these targets³ has had a detectable impact on the course of the epidemic in the United Kingdom. We conclude that it has now been brought under control, but it will be important to contain rapidly any new outbreaks

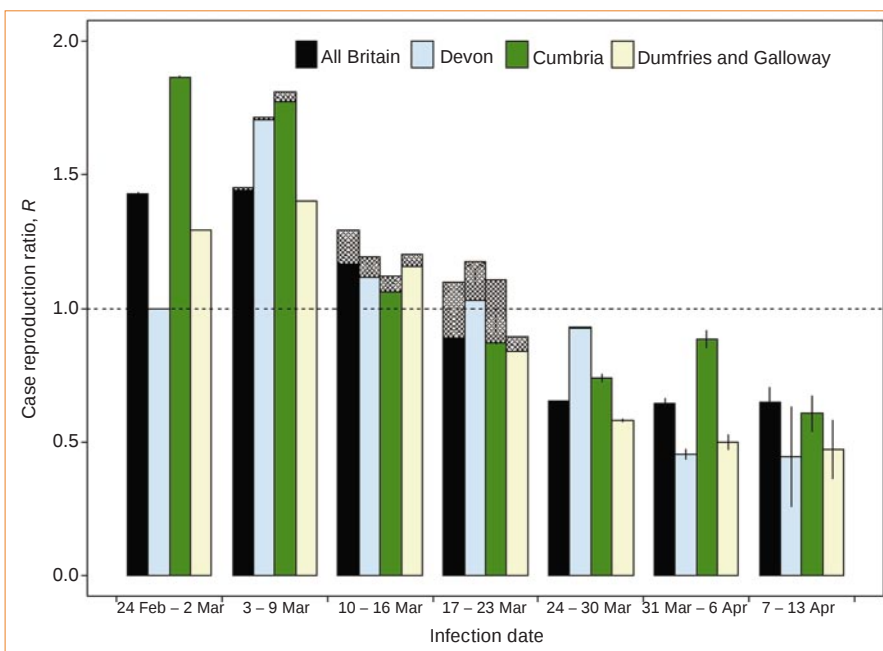


Figure 1 Changes through time in the case-reproduction ratio, R , for all of Britain and for the three most affected regions based on case report data up to 29 April. R values are for populations not samples. We estimated numbers of new cases not yet reported (censored) using the distribution of the interval between infection of a source case and reporting of cases infected from it, using negative binomially distributed variances to generate 95% confidence limits for R values corrected for censoring (vertical lines). We also adjusted R values to allow for the effect of culling incubating cases using the estimated proportion of total cases represented by at-risk farms culled (hatching). Incorrect infection dates – these could be earlier (rarely later) than estimated if clinical signs do not develop or are not reported – would lead to overestimation of recent R values. Incorrect infection sources would not affect average R values but could affect confidence limits.

in previously unaffected areas.

A key parameter describing the state of an epidemic is the case-reproduction ratio⁴, R , where R is the average number of new cases generated by each case (a case is defined here as the diagnosis of foot-and-mouth disease on a single farm). Only if R is less than 1 can the epidemic be regarded as ‘under control’⁵. We estimated R directly using a combination of demographic and contact tracing methods^{6–8}. The key inputs are MAFF data giving the estimated date of first infection (based on clinical assessment – age of oldest lesions and known incubation periods in different livestock species – and epidemiological information – proximity and date of infection of potential sources) and the exact location of each case diagnosed and reported¹. For some cases, MAFF identified the most probable source of infection using a combination of reported contacts with infected farms and (rarely) predicted patterns of aerial spread¹. We used the nearest infectious farm as the most probable source of infection for the remaining cases. We corrected R values for censoring (that is, new cases not yet reported) and allowed for the culling of at-risk farms (the ‘extended cull’; Fig. 1 legend).

The results indicate that R was initially high (3.3), reflecting extensive movements of infected livestock around the United Kingdom before restrictions were imposed on 23 February, after which R fell, but not

to below 1 (Fig. 1). Reduced average reporting-to-slaughter intervals (from 2.9 days in late February to 1.1 days in early April – together with reporting occurring an average of 0.7 days earlier), and intensification of the extended cull (from less than one additional farm culled per reported case before April to more than four during April) should have reduced R , but the effects would not have been immediate. However, the extended cull also had an immediate effect on reported incidence by removing incubating cases before they could be diagnosed. The true impact of increased culling effort was therefore not seen until cases generated by farms infected in late March were reported. Since that time, R has been below 1 for all Britain and for the three most affected regions (Fig. 1), implying that the epidemic is in decline. The expected rate of decline in incidence is related to the mean generation time (the average interval between infection of source and new cases, estimated as 7.6 days): for example, $R=0.65$ (the overall value in late March) corresponds to incidence halving every 12 days, consistent with subsequently reported trends.

The reduction in R and corresponding fall in the incidence of new cases was considerably greater than that recently predicted using a mathematical model⁹. Other factors that may have contributed include local ‘burn out’ of susceptible farms, more

susceptible farms being affected earlier than less susceptible farms⁸, and changes in underlying patterns of transmission.

Long-term projections are very sensitive to any further changes in R , and any intensification or relaxation of control efforts could greatly affect the final scale and duration of the epidemic. We note that the expected course of the epidemic is also very sensitive to the introduction of disease into previously unaffected areas and it is important that new outbreaks, especially in areas with high livestock densities, are contained rapidly and effectively.

Mark Woolhouse*, **Margo Chase-Topping***, **Daniel Haydon†**, **John Friar‡**, **Louise Matthews***, **Gareth Hughes*§**, **Darren Shaw***, **John Wilesmith||**, **Alex Donaldson§**, **Stephen Cornell#**, **Matthew Keeling#**, **Bryan Grenfell#**

*Centre for Tropical Veterinary Medicine, University of Edinburgh, Easter Bush, Roslin, Midlothian EH25 9RG, UK

†Department of Zoology, University of Guelph, Guelph, Ontario N1G 2W1, Canada

‡Eoscene Corporation, 700 5th Avenue, Seattle, Washington 98104, USA

§Institute for Animal Health Pirbright Laboratory, Ash Road, Pirbright, Surrey GU24 0NF, UK

||Veterinary Laboratories Agency, New Haw, Addlestone, Surrey KT15 3NB, UK

¶Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

#Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

1. Anon. *Vet. Rec.* **148**, 286–288 (2001).
2. Anon. *Vet. Rec.* **148**, 390–393 (2001).
3. <http://www.maff.gov.uk>
4. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Science, Oxford, 1991).
5. Woolhouse, M. & Donaldson, A. *Nature* **410**, 515–516 (2001).
6. Patil, G. P., Pielou, E. C. & Waters, W. E. (eds) *Statistical Ecology* Vol. II (Pennsylvania State Univ. Press, 1971).
7. Sellers, R. F., Garland, A. J. M., Donaldson, A. I. & Gloster, J. *Br. Vet. J.* **137**, 608–620 (1981).
8. Haydon, D. T., Woolhouse, M. E. J. & Kitching, R. P. *IMA J. Math. Appl. Med. Biol.* **14**, 1–9 (1997).
9. Ferguson, N., Donnelly, C. & Anderson, R. *Science* **13** April 2001 (<http://www.sciencexpress.org>).

Genomics

Annotation of the *Drosophila* genome

The publication of the genome of the fruitfly *Drosophila* by Celera Genomics last year generated considerable excitement¹. The initial release predicted 14,226 proteins, including about 550 that arise as the result of alternative splicing¹. Here we compare the annotation of the Celera *Drosophila* genome with known protein sequences lodged in a databank and find that although there is substantial agreement for many gene sequences, there are also numerous and significant discrepancies. These uncertainties call for caution in

interpreting predicted *Drosophila* genes.

Drosophila has been studied intensively and many of its genes have been well characterized experimentally. *Drosophila* protein sequences are collected in the SwissProt database and entries are updated periodically² to include information on alternatively spliced forms, alternative translation start sites, related motifs and functions of gene products.

We extracted 1,049 *Drosophila* protein sequences created before 1999 from the SwissProt database and compared them to the Celera proteome using BLAST³ (Table 1). We found that 26.2% of the SwissProt sequences perfectly matched a sequence from the Celera *Drosophila* genome (CDG) and that 28.8% were of identical length, with at least 99% similarity over the length of the SwissProt protein. The remaining 45% had sequence differences of more than 1%, including mismatches, insertions and deletions — small and large — spread over the protein's length. We list four representative examples here (for other examples, see <http://gnomic.stanford.edu>).

The HMCU homeobox protein encoded by the *cut* gene is 2,175 amino acids long (SwissProt database), but there is no significant BLASTP match in the CDG proteome. The closest match is to a protein (CP31015) that contains a BTB homodimerization domain (average probability of error, $E = 10^{-18}$). Even much lower E values do not automatically imply homology, as BLASTP is strictly a local alignment algorithm that matches subsegments of query and target sequences across the whole protein database. Moreover, no normalizations are used to account for proteins with different lengths and quality⁴. For HMCU, only 14% of the complete SwissProt protein is matched.

The CIK2 voltage-gated potassium channel (encoded by the *shaker* gene) has a SwissProt length of 643 amino acids and a CDG length of 708 (92% identity with CP23511). There is no alignment for the first 49 residues of the SwissProt sequence and the first 90 of the CDG sequence. The corresponding DNA sequence is in the Celera genome, suggesting that the CDG prediction could have missed an exon.

The product of the *trithorax* gene is 3,759 amino acids long according to the SwissProt database, but has 3,085 residues according to CDG (81% identity with CP25007). At residue 238 of the SwissProt protein there is an 11-residue discrepancy due to a deletion of the dinucleotide TC in the Celera DNA sequence, followed by a GG insertion 33 bases later, which restores the translation frame. The CDG sequence has two large deletions corresponding to 27 and 34 amino acids, respectively, at positions 450 and 1,998, and a 12-amino-acid mismatch at position 238.

Table 1 Comparison of matching SwissProt and CDG proteins

Identity	Same SP/CDG match length	SP/CDG match length within 1%	SP length < 99% CDG match	SP length >101% CDG match
100%	276 (26.3%)	32 (3.1%)	43 (4.1%)	27 (2.6%)
>99%	302 (28.8%)	107 (10.2%)	56 (5.3%)	22 (2.1%)
<99%	2 (0.2%)	11 (1.0%)	78 (7.4%)	93 (8.9%)
Total	580 (55.3%)	150 (14.3%)	177 (16.9%)	142 (13.5%)

Many other examples of protein-sequence misalignments are available at <http://gnomic.stanford.edu>. The CDG contains two identical copies of the male-specific doublesex protein (CP29316 and CP39510). Other exact duplication pairs include CP24170/CP24220 and CP40382/CP40384. We compared the CDG proteins to all possible translations of the genome sequence and found in every case that there was only one match in the genome. The duplicate protein copies are probably annotation errors. SP, SwissProt.

The neuronal-differentiation protein Prospero has 1,403 residues according to SwissProt and 1,703 according to CDG (CP38193). The SwissProt protein does not contain the first 300 residues of the CDG sequence, but the remainder matches perfectly. The extra DNA in the Celera transcript matches the SwissProt gene (EMBL messenger RNA Z11743) upstream of the translation-initiation codon, suggesting that one of the sequences incorrectly predicts the translation start site.

There are numerous examples of differences between SwissProt and CDG sequences in the length of amino-acid runs — for example, a glycine run in HMOC (orthodenticle homeotic protein), an asparagine run in HMCA (homeotic caudal protein) and a glutamine run in CEB (enhancer-binding protein).

These and many other differences between the SwissProt and CDG sequences seem to arise predominantly from annotation mistakes, such as omission of alternative splice forms from the CDG predictions. Some could also result from sequencing and assembly errors. True polymorphisms in *Drosophila* are surely present, but their frequency and extent are unknown. Alleles are identified for some genes in Flybase⁵, but many of these are deduced from targeted mutation experiments. The CDG annotation indicates that at least 1,600 genes disagree with their previously established sequences at the DNA level¹.

Gene-discovery algorithms exploit compositional differences between exons and introns, and look for signals such as splice sites and promoters. Searches using protein databases can help to locate genes that are similar to known genes but cannot identify completely new ones. Known protein-sequence properties could be used to improve exon prediction. Protein motifs and regular expression features such as zinc-finger motifs, ATP- and GTP-binding sequences, and motifs that are characteristic of kinase families are generally too small (5–15 residues) to be detected by BLAST, but may be used to support prediction of particular exons.

Another approach is to search for short 'words' (4–6 residues) that are significantly frequent⁶. Certain pentapeptides (such as GPPGP, CGKAF and HTGGK) occur fre-

quently in human proteins but are very rare in non-coding sequences, making them suitable as coding-region indicators. Other unusual features such as clusters of charged or hydrophobic residues and high-scoring potential transmembrane segments could also prove useful⁷. Statistically significant charge clusters of basic or acidic residues occur in roughly 19–24% of higher eukaryotic proteins⁸.

Although there are now more than 20 gene-prediction programs available, the task of annotation in eukaryotes is far from solved. Prediction solely on the basis of statistical and homology methods may prove to be intrinsically inadequate and experimental addenda may be needed. Results obtained by using expressed-sequence tags are valuable but tend to be biased by expression levels and are susceptible to contamination, fragmentation, fusion and other effects.

Proteomic studies using the present *Drosophila* genome sequence have significant limitations and the same will be true of the human genome. For now, these uncertainties should prescribe caution in interpreting newly predicted genes. Individual sequences should continually be corrected and refined by multiple rounds of annotation backed up with experimental data before the *Drosophila* genome can be considered complete and accurate.

Samuel Karlin, Aviv Bergman, Andrew J. Gentles

Mathematics Department, Stanford University, Serra Street, California 94305, USA
e-mail: karlin@math.stanford.edu

- Adams, M. D. et al. *Science* **287**, 2185–2195 (2000).
- Bairoch, A. & Apweiler, R. *Nucleic Acids Res.* **28**, 45–48 (2000).
- Altschul, S. F. et al. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
- Brochieri, L. & Karlin, S. J. *Mol. Biol.* **276**, 249–264 (1998).
- Flybase Consortium. *Nucleic Acids Res.* **27**, 85–88 (1999).
- Karlin, S. & Cardon, L. R. *Annu. Rev. Microbiol.* **48**, 619–654 (1994).
- Brendel, V. et al. *Proc. Natl Acad. Sci. USA* **89**, 2002–2006 (1992).
- Karlin, S. *Curr. Opin. Struct. Biol.* **5**, 360–371 (1995).

erratum

Flexible style that encourages outcrossing

Qing-Jun Li, Zai-Fu Xu, W. J. Kress, Yong-Mei Xia, Ling Zhang, Xiao-Bao Deng, Jiang-Yun Gao, Zhi-Lin Bai
Nature **410**, 432 (2001)

In the legend to Fig. 2, the full line refers to the cataflexistyle flower and the dotted line to the hyperflexistyle flower (and not vice versa, as published).

A glia-derived acetylcholine-binding protein that modulates synaptic transmission

August B. Smit*, Naweed I. Syed†, Dick Schaap‡§, Jan van Minnen*, Judith Klumperman||, Karel S. Kits*, Hans Lodder*, Roel C. van der Schors*, René van Elk*, Bertram Sorgedraeger*, Katjuša Brejc¶, Titia K. Sixma¶ & Wijnand P. M. Geraerts*

* Department of Molecular and Cellular Neurobiology, Research Institute Neurosciences Vrije Universiteit, Faculty of Biology, De Boelelaan 1087, 1081 HV Amsterdam, The Netherlands

¶ Division of Molecular Carcinogenesis, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

† Department of Cell Biology and Anatomy, University of Calgary, T2N4N1, Calgary, Canada

|| Department of Cell Biology, University Medical Center, Institute of Biomembranes and Center for Biogenetics, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands

‡ Organon Teknika B.V., Bioscience Research Unit, Boseind 15, Boxtel, The Netherlands

There is accumulating evidence that glial cells actively modulate neuronal synaptic transmission. We identified a glia-derived soluble acetylcholine-binding protein (AChBP), which is a naturally occurring analogue of the ligand-binding domains of the nicotinic acetylcholine receptors (nAChRs). Like the nAChRs, it assembles into a homopentamer with ligand-binding characteristics that are typical for a nicotinic receptor; unlike the nAChRs, however, it lacks the domains to form a transmembrane ion channel. Presynaptic release of acetylcholine induces the secretion of AChBP through the glial secretory pathway. We describe a molecular and cellular mechanism by which glial cells release AChBP in the synaptic cleft, and propose a model for how they actively regulate cholinergic transmission between neurons in the central nervous system.

Chemical synaptic transmission is an important mode of signalling in the nervous system. The prevailing view is that chemical synaptic transmission exclusively involves bipartite synapses consisting of pre- and postsynaptic elements and a synaptic cleft, in which presynaptically released transmitter binds to cognate receptors in the postsynaptic cell. However, there is accumulating evidence that synapses have a more complex organization, and that glial cells not only actively participate in but can also modulate synaptic transmission^{1,2}.

For example, several studies have shown that presynaptically released glutamate not only binds receptors on the postsynaptic neuron^{3,4}, but may also activate AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid)-type glutamate receptors on perisynaptic astrocytes^{5–9}, causing a rapid increase in intracellular Ca^{2+} concentration¹⁰ and Ca^{2+} -dependent release of glutamate from glia^{5,9,11,12}. Glutamate released by the perisynaptic astrocytes in turn activates NMDA (*N*-methyl-D-aspartate)-type receptors on the presynaptic neuron, thereby increasing the release of glutamate and enhancing synaptic transmission^{5,13,14}. Similar modulatory mechanisms involving perisynaptic glial cells have been suggested for GABA (γ -aminobutyric acid)-mediated synapses, where GABA released by astrocytes potentiates inhibitory synaptic transmission¹⁵, and for the cholinergic neuromuscular junction, where acetylcholine (ACh) released by activated perisynaptic glia^{16,17} probably feeds back onto the presynaptic nerve terminal¹⁸. Thus, for various types of synapse the glial cells complete a tripartite configuration, in which they can feed back modulatory signals to neuronal synaptic elements by releasing endogenous transmitter. As such, glial and neuronal cells can form integral modulatory components of synaptic function¹⁹.

In this study we explored the existence of alternative modulatory mechanisms, in which glial cells at tripartite synapses would interact directly with synaptic transmitters. We found that perisynaptic glial

cells of a molluscan cholinergic synapse respond to ACh by releasing a soluble ACh-receptor-like protein²⁰ into the synaptic cleft. This protein captures presynaptically released ACh and suppresses synaptic transmission. The release of this decoy glial receptor protein into the synaptic cleft provides a novel mechanism by which glial cells can modulate the efficacy of neuronal transmission in the central nervous system (CNS).

Glia modulate cholinergic synaptic transmission

To study the modulatory role of glia in interneuronal synaptic transmission, we cultured identified neurons from the CNS of the mollusc *Lymnaea stagnalis* with or without glial cells. Synapses between a presynaptic cholinergic neuron and its postsynaptic partner, and between a presynaptic dopaminergic neuron and its postsynaptic target, were reconstructed^{21–23}. Glial cells isolated from the CNS and plated close to the synapses migrated to the synapse within an hour. In the absence of glial cells, a train of induced action potentials in the cholinergic presynaptic neuron produced facilitatory excitatory postsynaptic potentials (EPSPs) in the postsynaptic cell, which often led to spiking activity (Fig. 1a). In the presence of glial cells, however, the presynaptically induced action potentials failed to elicit facilitatory EPSPs and action potentials in the postsynaptic cell (Fig. 1b). The glial cells did not affect dopaminergic synaptic transmission (Fig. 1c, d). These data show specific modulation by glial cells of the efficacy of cholinergic synaptic transmission, which is apparent only after high-frequency stimulation of the presynaptic cell. When neurons were cultured in a triplet configuration in which the presynaptic neuron forms two synapses with postsynaptic partners, only the synapse co-cultured with glia showed synaptic depression (Fig. 1f), leaving the non-glia-bearing synapse unaffected (Fig. 1e). Thus, the presence of glial cells at one synaptic site does not render the transmitter release machinery ineffective at other synapses.

To test the glial cell's ability to sense presynaptic transmitter release, we placed an isolated glial cell close to the presynaptic cholinergic neuron (Fig. 2a). Depolarization-induced transmitter

§ Present address: Intervet, de Körverstraat 35, 5830 AA Boxmeer, The Netherlands.

release from the presynaptic neuron was detected in the glial cell as non-synaptic depolarizing potentials. To define the transmitter-induced response in the glial cell in more detail, we recorded responses from an isolated glial cell to exogenously applied ACh under whole-cell voltage clamp conditions. Pressure application of 10 μ M ACh onto glial cells evoked a rapid inward current (Fig. 2b) with a median effective concentration (EC_{50}) of 2.4 μ M (Fig. 2d). These effects were mimicked by nicotine, but the nicotine-induced current inactivated much faster than the ACh-induced current (Fig. 2c). To define the nature of the receptor subtype that mediates the glial cholinergic response, we tested the effect of α -bungarotoxin (α -Bgt, a specific blocker of nAChRs of the vertebrate neuromuscular junction and of the homopentameric $\alpha 7$ -type nAChR²⁴). In the presence of α -Bgt the ACh-induced inward current was almost completely suppressed (Fig. 2e), indicating that the cholinergic

response in the glial cells may be mediated by an ionotropic, α -Bgt-sensitive nAChR. This glial nAChR has a reversal potential of 0 mV and conducts sodium ions, and is different from the chloride channels reported previously in molluscan neurons^{25,26}.

The physiological data led us to hypothesize that glial-cell-induced suppression of cholinergic synaptic transmission involves release of ACh from the presynaptic neuron, which depolarizes the glial cell through activation of an α -Bgt sensitive nAChR. The ACh-induced depolarization might in turn cause the release of an unknown factor, which suppresses synaptic transmission.

Glia express an nAChR-like protein

To test the above hypothesis, we first used rhodamine-conjugated α -Bgt to determine whether the $\alpha 7$ -type of nAChR was present on the glial cells. The α -Bgt labelling in sections of the *Lymnaea* CNS was intense, and in particular localized to glial cells (see below). Glial cells stained in culture showed a small amount of labelling of the cell surface (not shown); however, after permeabilization almost all labelling was punctate and localized to the cytoplasm of the glial cells (Fig. 3a). As the labelled unknown cytoplasmic factor probably contained an acetylcholine-binding site, we reasoned that it might be involved in the modulation of cholinergic transmission. To identify this factor, we used α -Bgt-biotin conjugated to streptavidin to

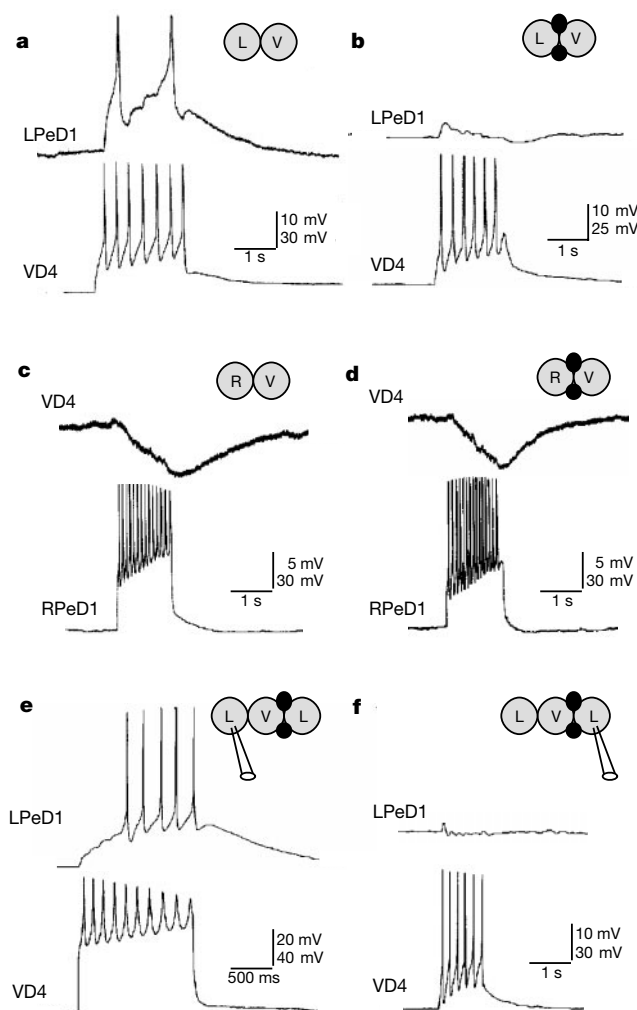


Figure 1 Glial cells cultured with synaptically paired neurons specifically inhibit cholinergic synaptic transmission. Synapse pairs represented schematically (V, VD4; L, LPeD1; R, RPeD1). Black circles, glia. **a**, VD4–LPeD1 pairs form a cholinergic excitatory synapse. Action potentials induced in the presynaptic cholinergic neuron VD4 generate EPSPs and action potentials in LPeD1 ($n = 13$ cell pairs). **b**, As in **a**, but cultured with glial cells that suppress synaptic transmission ($n = 13$). Residual response can be blocked by hexamethonium and is mediated by a nAChR. **c**, RPeD1–VD4 pairs form an inhibitory dopaminergic synapse. **d**, Co-cultured glia do not suppress dopaminergic synaptic inhibition. **e, f**, Dual-synapse configuration (two cholinergic synapses on a single presynaptic cell). **e**, In the synapse without co-cultured glia, VD4 action potentials lead to normal synaptic transmission with EPSPs and action potentials in LPeD1 ($n = 5$). **f**, In the synapse with co-cultured glia, synaptic transmission is suppressed ($n = 7$).

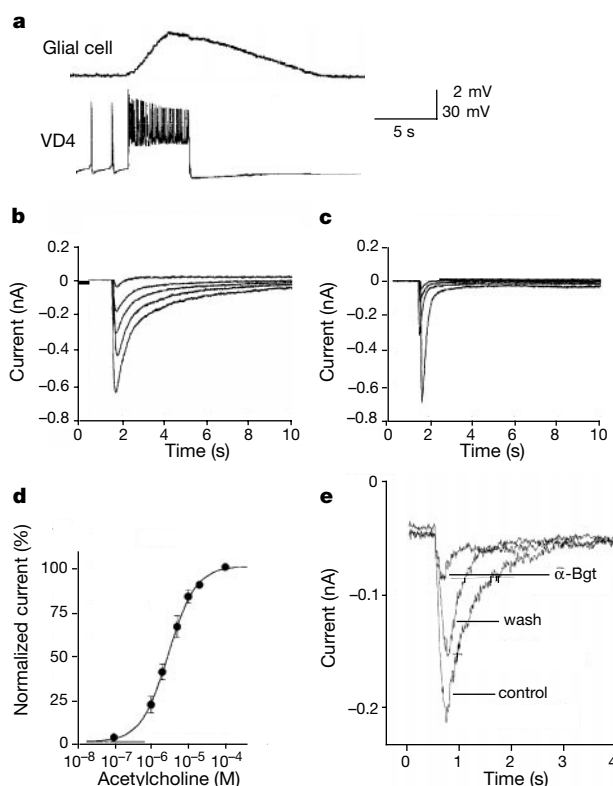


Figure 2 Glial cells detect ACh through a nicotinic ACh receptor. **a**, Glial cells held close to a VD4 neuron respond to action-potential-induced transmitter release with a compound depolarizing potential ($n = 5$). **b**, Whole-cell voltage clamp responses from an isolated glial cell to 10 μ M ACh applied by pressure ejection, recorded at holding potentials ranging from -100 mV (largest response) to -20 mV (smallest response; step size 20 mV). **c**, As in **b**, but showing responses to nicotine. **d**, Dose–response curve of glial cells to ACh (applied by pressure ejection). Responses were normalized to the maximum amplitude in the ACh experiment. Hill coefficient, 1.94. EC_{50} from fitted line is 2.4 μ M. Recorded at a holding potential of -80 mV; $n = 5$ cells; intervals between applications 50 s, sufficient to allow complete recovery from desensitization. **e**, Block of the current response to ACh (applied iontophoretically) by α -Bgt (10 μ M, applied by pressure ejection). Holding potential, -80 mV.

specifically pull down α -Bgt-binding proteins in particulate and soluble fractions of homogenates of the *Lymnaea* CNS. In contrast to subunits of nAChRs, which typically have relative molecular masses above 50,000 ($M_r > 50K$), SDS-polyacrylamide gel electrophoresis revealed a protein with an M_r of only around 25K, specifically enriched in the soluble fraction of the CNS homogenate (Fig. 3b). A partial amino-terminal amino-acid sequence, LDRA-DILYNIRQ, was obtained by western blotting and Edman degradation. On the basis of this, we generated a polymerase chain reaction (PCR) product on a complementary DNA library of the *Lymnaea* CNS and used it to screen the library. Sequence analysis of a 980-base-pair (bp) cDNA revealed a single open reading frame encoding a protein of 229 amino acids (Fig. 3c). The length of this cDNA corresponds well to the roughly 1,000-nucleotide transcript determined by northern blotting (Fig. 3d).

Because the α -Bgt-binding protein isolated from the CNS starts with Leu at position 20, the mature protein is only 210 amino acids long. In the precursor it is preceded by a signal sequence of 19 amino acids. To determine its relative molecular mass, the protein pulled down by α -Bgt was purified by high-performance liquid chromatography (HPLC), coupled on-line to a mass spectrometer (Q-TOF (quadrupole time-of-flight)) and shown to consist of a single molecular species of M_r 24,720.4 (not shown). Deglycosylation yields a protein with an M_r of only $23,832.9 \pm 2.8$ (not shown), corresponding to the mass of the mature protein predicted from the cDNA of 23,834. Thus, the α -Bgt-binding protein is a $\sim 23.8K$ glycoprotein with a glycosyl group of $M_r \sim 888$.

The α -Bgt-binding protein has moderate sequence identity with subunits of the Cys-loop family of ligand-gated ion channels (the nicotinic acetylcholine receptors (nAChRs), GABA_A, GABA_C, glycine and 5HT₃ receptors^{27–32}). These receptors comprise five

subunits, each consisting of an N-terminal extracellular domain, which is involved in ligand-binding, and a carboxy-terminal half containing four transmembrane segments (M1–M4), which make up the channel pore. The α -Bgt-binding protein is much shorter than these subunits and aligns only with the extracellular part, lacking pore-forming transmembrane regions. No homologous proteins of this size were found in sequence databases.

The highest sequence identity of AChBP is with the extracellular domains of the α -subunits of the nAChRs (Fig. 3c), and notably also with the $\alpha 7$ nAChR (23.9% in 210 amino acids). We therefore tentatively designated the α -Bgt-binding protein as acetylcholine-binding protein (AChBP). Although the sequence of AChBP is clearly related to those of the nAChRs, it is not a truncated form of a nAChR α -subunit. First, the AChBP transcript, at 1,050 bases, is too small to accommodate a full-length α -subunit of a *Lymnaea* nAChR (typically more than 1,800 nucleotides). We did not find alternatively spliced transcripts encoding a sequence-related nAChR. Second, rescreening the CNS library (at low stringency) using AChBP cDNA did not reveal any nAChR-related cDNAs among 50 sequenced. Third, as shown by affinity isolation and mass determination, the AChBP of 23.8K as predicted from the cDNA is present in the CNS. In addition, immunoprecipitation using an AChBP-specific antibody did not indicate the existence of larger proteins with AChBP epitopes.

In the nAChRs, both α - and non- α -subunits (or only α -subunits in the case of the $\alpha 7$ nAChR) contribute to ligand binding³², with the α -subunit contributing the principal components of the binding site. In the nAChR, specific residues residing in three loops are involved in ACh binding (for example, Trp 86, Tyr 93, Trp 149, Tyr 151, Tyr 190, the Cys 192, 193 doublet and Tyr 198 (ref. 32)). Except for Tyr 151, these residues are all conserved in AChBP

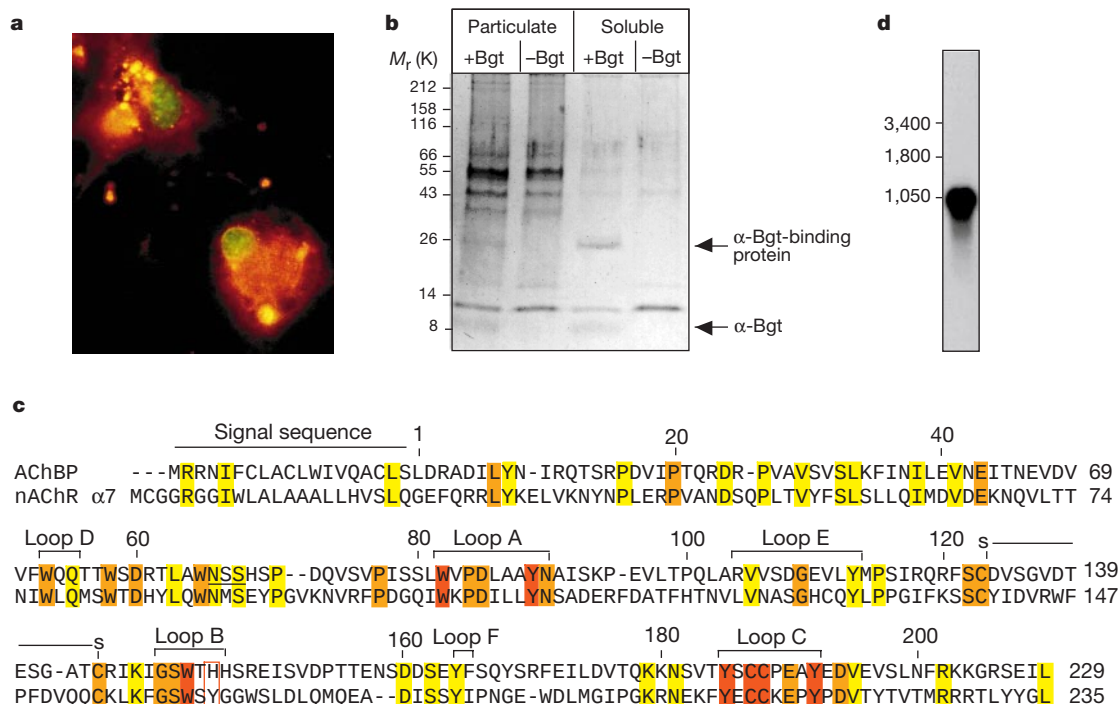


Figure 3 Characterization of α -bungarotoxin-binding protein from the *Lymnaea* CNS. **a**, α -Bgt-rhodamine labelling of glial cells freshly isolated from the CNS and after permeabilization. Cytoplasmic α -Bgt-rhodamine labelling is yellow/orange; nuclei counterstained with Syto-16 (green). **b**, Affinity isolation of α -Bgt-binding proteins from particulate and soluble fractions of 40 *Lymnaea* CNSs, with or without α -Bgt and run on 8% SDS-PAGE. An abundantly isolated 25K α -Bgt binding protein is indicated. **c**, Amino acid sequence of AChBP and comparison with ligand-binding subunit of the $\alpha 7$ nAChR.

Three ligand-binding loops A, B and C (as described for the nAChRs), and loops D, E and F of the complementary binding site, are indicated. Red, residues previously shown to be involved in ligand binding; orange, residues conserved in the nAChR family; yellow, residues conserved between AChBP and $\alpha 7$ nAChR. Underlined, consensus glycosylation site. **d**, Northern blotting at low stringency indicates the existence of a single AChBP transcript (numbers indicate bases).

(Fig. 3c). The characteristically conserved cysteine disulphide bridge forming a 15-residue loop in the extracellular domain of the ligand-binding subunits of nAChRs is present at position 123–136 in AChBP (Fig. 3c). However, the cysteines are spaced by only 12 instead of 13 amino acids, and sequence conservation of the loop residues is lost. The sequence similarity between AChBP and the nAChR α -subunits, and in particular the conservation of the ligand-binding residues, led us to test the ligand-binding characteristics of AChBP with nAChR ligands.

Multimerization and ligand binding by AChBP

To test ligand binding by AChBP, we expressed the protein in yeast²⁰ (recAChBP). When tested under native conditions on a gel-permeation column it runs at an apparent M_r of ~ 160 K (Fig. 4a), indicating that it forms a (soluble) multimer of 5 or 6 subunits (X-ray crystallography²⁰ showed that AChBP was a pentamer). Subunits of the nAChRs form homo- and heteropentamers, both *in vivo* and *in vitro*^{33,34}. In the $\alpha 7$ nAChR the determinants for pentamerization have been mapped to regions in both the extracellular domain³⁵ and the first transmembrane domain³⁶. In AChBP, which is analogous to the extracellular domain of nAChRs, this domain is apparently sufficient to determine pentamerization. Studies on the $\alpha 7$ nAChR have indicated the importance of the M1 region in the expression of the extracellular domain³⁷.

To determine the ligand-binding characteristics of the AChBP multimer, we performed a series of binding assays. First, we determined the binding curve of α -Bgt to recAChBP, and calculated an affinity of 3.5 nM (Fig. 4b). Using 125 I- α -Bgt in a competitive binding assay we then tested ligands of several types of ligand-gated ion channel on recAChBP (ACh, serotonin, GABA, glycine and

glutamate). Both ACh and serotonin competed with 125 I- α -Bgt binding at 4.2 μ M and 269 μ M, respectively, whereas GABA, glycine and glutamate did not (Fig. 4c). AChBP binds ACh but does not contain esterase activity (not shown). Second, we studied the ligand-binding characteristics of AChBP in more detail using agonists (Fig. 4d) and antagonists (Fig. 4e) of the AChRs. Nicotine, the typical agonist of nAChRs, is a high-affinity ligand of recAChBP (IC_{50} 98 nM). The apparent affinity of ACh for AChBP is 42-fold lower than that of nicotine, a typical feature of the $\alpha 7$ nAChR³⁸. In addition, the affinities for both ligands are much lower (ACh, 1,000-fold; nicotine, 18-fold) than observed for the high-affinity heteromeric nAChRs (for example, the $\alpha 4\beta 2$ subtype). Epibatidine, a high-affinity agonist of nAChRs³⁹, also binds with high affinity to recAChBP (IC_{50} 1.4 nM). Other cholinergic agonists bind with low affinity: for example, carbachol and choline have IC_{50} s of 43 and 190 μ M, respectively.

Typical antagonists of the nAChRs, such as tubocurarine and α -Bgt, have a high affinity for AChBP (IC_{50} s of 93 nM and 3.5 nM, respectively). The cholinergic antagonist succinylcholine has a very low affinity for AChBP (IC_{50} 7.9 mM). Muscarinic receptor antagonists also bind to recAChBP with relatively high affinities: for example, the muscarinic allosteric modulator gallamine (IC_{50} 39 nM) and the muscarinic antagonist atropine (IC_{50} 5.3 μ M). Finally, bipinnatin-B, a synthetic form of coral toxin and a general blocker of nAChRs⁴⁰, covalently blocked recAChBP (see Supplementary Information). The binding characteristics of recAChBP are summarized in Fig. 4f. In contrast to the $\alpha 7$ nAChR⁴¹ (Hill coefficient for ACh ~ 1.4), AChBP has Hill coefficients of less than 1 for all agonists and some antagonists, which might represent true negative cooperativity or some micro-heterogeneity due to

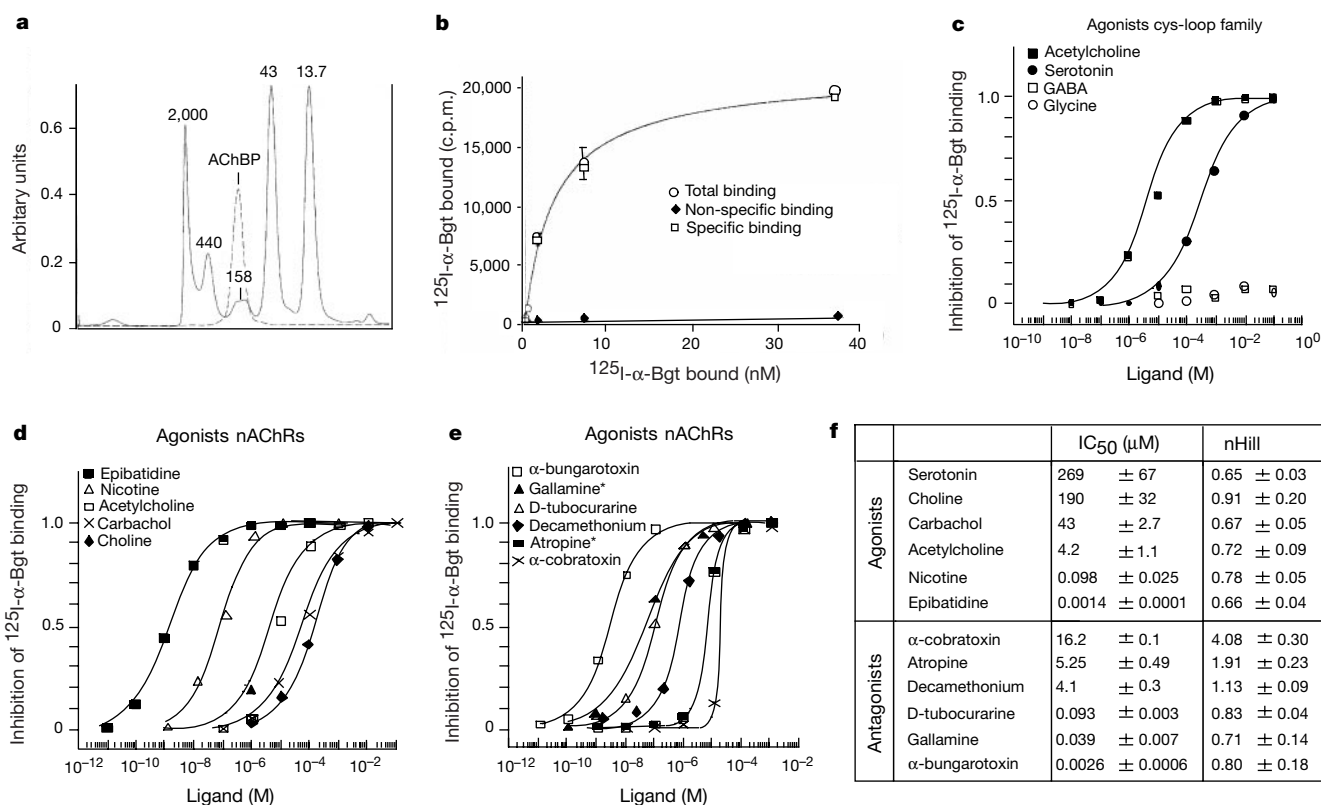


Figure 4 Ligand-binding characteristics of AChBP. **a**, The running profile of recAChBP under native conditions. Marker proteins and AChBP were run separately and chromatograms were superimposed. M_r as indicated. **b**, Saturation curve of 125 I- α -Bgt to AChBP. **c**, Various ligands, ACh, serotonin, glycine, GABA, and glutamate, were incubated with His-tagged AChBP and allowed to compete for binding with 125 I- α -Bgt. Binding

curves for ACh and serotonin and IC_{50} values are indicated. GABA, glycine, and glutamate do not show competition binding. **d**, **e**, Competition binding on AChBP as in **c** with agonists and antagonists of the nAChRs. **f**, Summary of IC_{50} values and Hill coefficients (nHill) of the plots in **c**–**e**. Note that gallamine and atropine have mixed pharmacology and are also antagonists of the muscarinic AChRs.

either non-equivalence of the subunits or slightly different forms of symmetric oligomers. Thus, AChBP shows a mixed nicotinic–muscarinic pharmacology with many features of nAChRs of the $\alpha 7$ - and $\alpha 9$ -subtypes^{42,43}.

Cellular and subcellular localization of AChBP

Localization studies of AChBP in the *Lymnaea* CNS, tested by labelling of CNS sections with rhodamine- α -Bgt, immunocytochemistry and *in situ* hybridization of AChBP, show co-expression of the AChBP gene and protein in glial cells (Fig. 5a–c). The AChBP-expressing glial cells are localized between neuronal cell bodies (Fig. 5d, e) and are also densely packed between axon terminals (Fig. 5c, g). When stained in culture, only about 50% of the glial cells produce AChBP.

To determine the cellular route for the secretion of AChBP in the glial cells, we performed immunogold labelling of AChBP on ultrathin cryosections of the *Lymnaea* CNS. AChBP protein was readily visualized in the early secretory pathway of the glial cells, for example in the endoplasmic reticulum and the Golgi apparatus (Fig. 5f). In addition, it was found in denser granule-like compartments that occurred near the Golgi apparatus (Fig. 5f) and in the long processes of the cells (Fig. 5g). The morphology, subcellular distribution and high content of AChBP in these dense granules

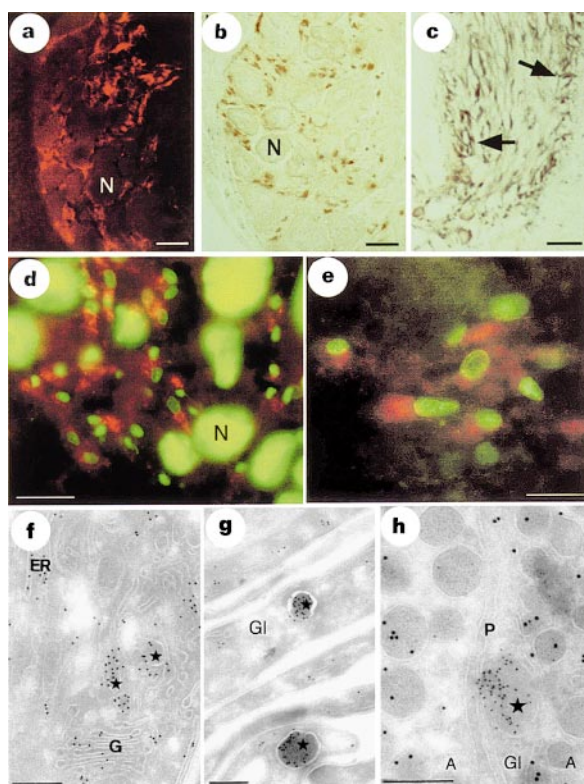


Figure 5 Cellular and subcellular expression of AChBP. **a**, Rhodamine- α -Bgt labelling of the cerebral commissure of the *Lymnaea* CNS, showing AChBP-containing glial cells. **b**, Immunostaining of glial cells in the cerebral ganglion. **c**, *In situ* hybridization showing glial cells between axon endings of egg-laying hormone neurons in cerebral commissure (arrows). Scale bar, 100 μ m; N, neuron. **d**, **e**, Rhodamine- α -Bgt labelling of glia (punctate labelling) in the cerebral ganglia. Neurons are devoid of signal. Scale bar, 100 μ m (**d**), 30 μ m (**e**). **f–h**, Ultrathin cryosections of the *Lymnaea* CNS, immunogold labelled for AChBP (10 nm gold) or double-labelled for AChBP (5 nm gold) and egg-laying hormone (10 nm gold). **f**, Anti-AChBP densely labels the endoplasmic reticulum (ER) and Golgi apparatus (G) of a glial cell. Stars, LDCVs in AChBP-containing glia. **g**, AChBP LDCVs in glial processes (Gl). **h**, A glial cell process (Gl) containing an LDCV with AChBP (star) between axons (A) of egg-laying hormone neurons, with hormone-containing vesicles. P, plasma membrane; bars in **f–h**, 200 nm.

indicate that they probably represent large, dense-core secretory vesicles (LDCVs) involved in AChBP secretion.

We then focused on neuroendocrine cells that produce the egg-laying hormone, because their electrical activity is known to be regulated by ACh. Double-immunolabelling for AChBP and a peptide of the egg-laying hormone precursor showed AChBP-containing LDCVs in glial cell processes that were intermingled with the axon terminals of cells producing egg-laying hormone (Fig. 5h).

Release of AChBP from glia is mediated by an nAChR

Because AChBP is produced in the secretory pathway of glial cells and binds acetylcholine, and because ACh depolarizes glial cells, we directly assessed whether ACh also induces the cellular release of AChBP. After metabolic labelling of glial cells *in vitro* with ³⁵S-labelled cysteine/methionine, we quantified the release of AChBP using immunoprecipitation (Fig. 6a). ACh (10 μ M) applied to glial cells *in vitro* induced AChBP release (Fig. 6b). The induction of AChBP release by ACh could be blocked almost completely by α -Bgt (Fig. 6b), which indicates that the glial nAChR is involved in the release of AChBP.

Suppression of cholinergic transmission by AChBP

Because the glial cells suppress cholinergic transmission, produce AChBP and release it in response to ACh, it is important to show that AChBP can suppress synaptic transmission. To test this, we reconstructed the cholinergic synapse *in vitro* and tested synaptic transmission in the absence (Fig. 7a) and presence (Fig. 7b) of recAChBP (10 μ M). We found that recAChBP significantly suppressed synaptic transmission, in a manner similar to that observed with co-cultured glial cells (Fig. 1b). The first EPSP was largely unperturbed, whereas subsequent synaptic potentials were completely or significantly suppressed by recAChBP. Bipinnatin-B-inactivated AChBP, on the other hand, failed to suppress synaptic transmission between the paired cells (Fig. 7c). In this set-up, bipinnatin-B left intact the response of the postsynaptic receptors (Fig. 7d). Finally, to test whether in the neuron–glia co-culture it is indeed endogenous AChBP that suppresses synaptic transmission, we used α -Bgt to block binding of ACh to AChBP and to stop the release of AChBP mediated by the glial nAChR. Consistent with our hypothesis, addition of α -Bgt to the culture removed the glial (AChBP)-induced suppression of synaptic transmission between the paired cells (Fig. 7e, f). Our data show that AChBP can capture presynaptically released ACh and thus act as a synaptic ACh buffer, causing immediate suppression of synaptic transmission.

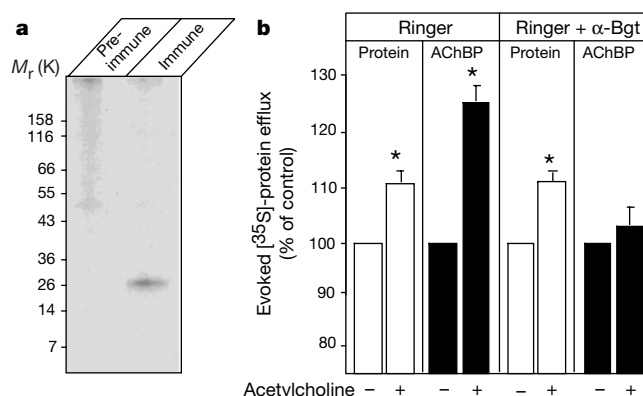


Figure 6 Glial cells release AChBP. **a**, Immunoprecipitation of radiolabelled AChBP from the *Lymnaea* CNS using a polyclonal antibody. **b**, Induced release of radiolabelled AChBP or other cellular proteins by ACh, quantified after (immuno)precipitation. Release of AChBP is almost completely blocked by pre-incubating glia in α -Bgt. Asterisk, significantly different from control ($P < 0.05$).

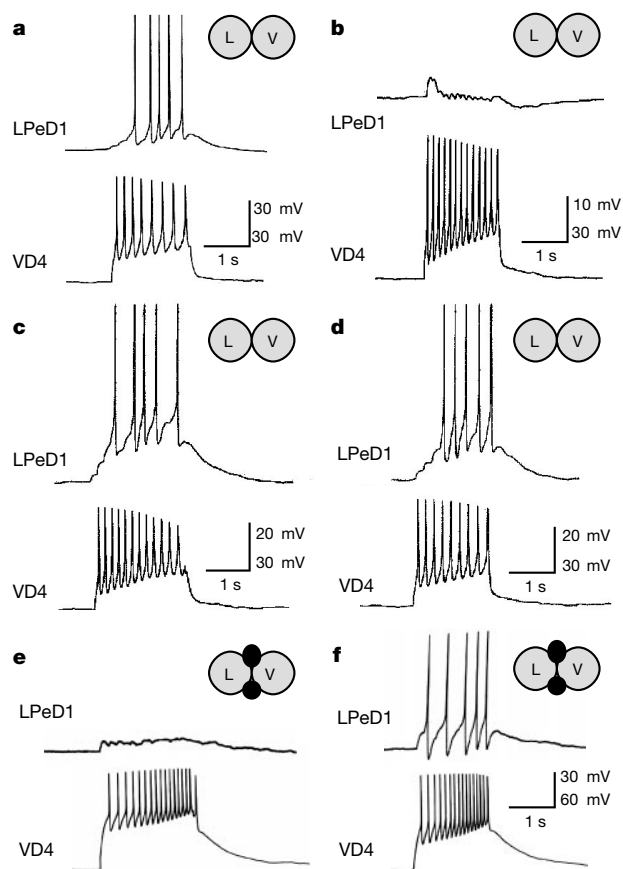


Figure 7 AChBP suppresses cholinergic neurotransmission. **a**, Normal cholinergic excitatory synaptic transmission between VD4 and LPeD1. **b**, Synaptic transmission is suppressed by superfusion with 10 μ M AChBP. **c**, Synaptic transmission is not suppressed by superfusion with biotinylated AChBP. **d**, Biotinylated AChBP superfusion alone does not suppress synaptic transmission. **e**, Suppression of transmission in the glia–neuron co-culture (VD4–LPeD1–glia) is lifted (**f**) by the addition of α -bungarotoxin. Symbols as in Fig. 1.

Discussion

Using a combined molecular, pharmacological and physiological approach we have identified a soluble transmitter receptor, AChBP, and shown that it suppresses cholinergic synaptic transmission. Its inclusion in the secretory pathway of glial cells and its stimulus-dependent release into the synaptic cleft constitute a novel molecular and cellular mechanism that is sufficient to account for its modulatory role in synaptic transmission.

On the basis of our *in vitro* data (Fig. 6), we suggest that there is a basal level of AChBP in the synaptic cleft, maintained by continuous release from the synaptic glial cells. Under conditions of active presynaptic transmitter release, high millimolar concentrations of free ACh⁴⁷ will probably activate both postsynaptic receptors and the nAChRs on the synaptic glial cells (EC_{50} in the micromolar range; Fig. 2d), which would enhance the release of AChBP, thus increasing its concentration in the synaptic cleft (Fig. 8). This may either diminish or terminate the ongoing ACh response, or raise the concentration of basal AChBP to the extent that subsequent responses to ACh are decreased. Accurate models of the kinetics of AChBP in synaptic modulation will require more data, such as concentrations of ACh, AChE and postsynaptic nAChRs, the localization and individual affinities of these components and the diffusion characteristics of ACh^{44–46}. Similarly, they will have to incorporate the effect of the ultrastructure of the synapse that

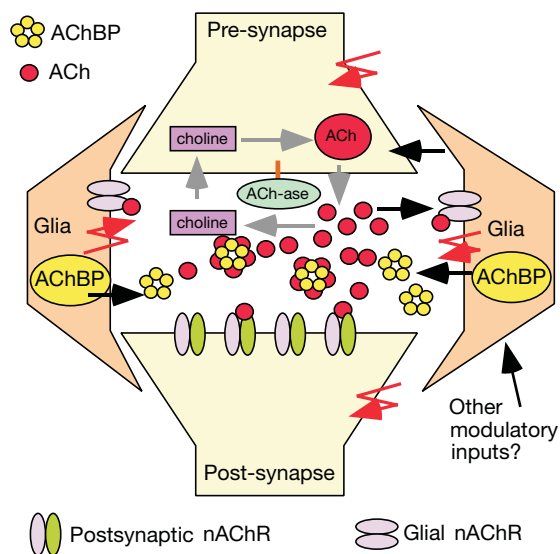


Figure 8 Model of the role of AChBP in neurotransmission. A basal level of AChBP is present in the synaptic cleft. Presynaptic ACh release can lead to activation of postsynaptic receptors and to EPSPs. In parallel, nAChRs on glia are activated, causing increased release of AChBP into the synapse, which leads to suppression of cholinergic transmission (see Discussion).

determines the kinetic properties of these constituents (that is, small processes of synapse-invaginating glial cells that locally release AChBP).

However, for the sake of simplicity one could consider two extreme scenarios of cholinergic transmission—under high and low concentrations of AChBP. On the basis of the measured dissociation constant of 4 μ M, a concentration of 80 μ M AChBP and 80 μ M ACh would generate a free ACh concentration of 16 μ M, which is probably sufficient to activate postsynaptic receptors. At high AChBP concentrations (for example, 400 μ M), however, the concentration of free ACh would drop to 0.9 μ M, which is likely to be insufficient to activate postsynaptic receptors. Moreover, the activation of postsynaptic nAChRs at concentrations below their EC_{50} value for ACh drops with the square of the ACh concentration⁴⁴, thus limiting further nAChR activation at low ACh concentrations. So, if AChBP release would bring down the free ACh concentration to values below the EC_{50} , the postsynaptic response would be substantially affected. As such, the concentration of AChBP in the synapse might critically determine whether transmission is fully active or suppressed.

When ACh is bound to AChBP in the synaptic cleft, acetylcholinesterase will hydrolyse only the free ACh, thereby keeping the ACh concentration minimal and draining it from the AChBP buffer. Under this condition, ACh bound to AChBP is unable to activate postsynaptic nAChRs, while the concentration of remaining free transmitter might not be sufficient to do so.

In our experimental set-up AChBP leads to complete, or nearly complete (Fig. 7e), suppression of synaptic transmission. *In vivo*, however, synaptic glial cells may be less numerous, exhibit different thresholds for AChBP release or even have different morphology in the synaptic cleft, hence leading to a less stringent AChBP block and a more balanced modulation of the efficacy of synaptic transmission. Moreover, *in vivo*, synaptic glial cells might integrate various signals and even at inactive synapses tune the rate of AChBP release (up or down), thereby setting the basal synaptic level of AChBP on which modulation of synaptic efficacy will occur during active transmission (Fig. 8). In addition to a regulated release of AChBP, its removal from the synaptic cleft might also be actively controlled to determine its precise concentration in the synaptic cleft. □

Methods

Cell culture

Ganglia were isolated and prepared for dissection⁴⁸. Identified neurons were isolated from the CNS, plated on poly-L-lysine-coated dishes containing brain-conditioned medium, and paired in a soma–soma configuration as described²¹. Glial cells were extracted from pedal ganglia using a glass pipette (30–40 µm diameter) and gentle pressure suction. Isolated glia were then gently flushed close to the soma–soma paired neurons. Identified neurons used^{21,22} were: VD4, visceral–dorsal 4 neuron; LPeD1, left pedal–dorsal 1 neuron; RPeD1, right pedal–dorsal 1 neuron.

Electrophysiology

For intracellular recordings, conventional sharp electrode recording techniques were used⁴⁸. Whole-cell voltage clamp recordings were made at room temperature from freshly dissociated glial cells in saline containing (in mM) 30 NaCl, 1.7 KCl, 10 NaH₂SO₄, 1.5 MgCl₂, 4 CaCl₂, 5 NaHCO₃ and 10 HEPES, pH 7.8, with NaOH, using an EPC7 amplifier (List) and pipettes filled with saline containing (in mM) 29 KCl, 2 MgATP, 2.3 CaCl₂, 11 EGTA, 10 HEPES, pH 7.4, with KOH. After reaching whole-cell configuration, cell capacitance was compensated (average 4.95 ± 0.97 pF); series resistance was not compensated. ACh, nicotine and α-Bgt were applied by pressure application.

Pharmacology

For competition experiments, we used 0.2 nM AChBP in assay buffer (PBS; Tris, 20 mM; BSA, 1 mg ml⁻¹; pH 8.0). Ligands or buffer were added and incubated for 15 min at 20 °C. The final concentration of [¹²⁵I]-α-Bgt was 0.1 nM. After incubation with continuous shaking for 1 h at 20 °C, we added Co²⁺ bead slurry (Clontech) pre-washed with assay buffer. After 15 min the beads were washed, centrifuged and scintillation counted. Experiments were performed fourfold. Curves were fitted to the data.

PCR and cDNA library screening

A degenerate oligonucleotide was synthesized on the basis of residues 1–10 of AChBP (LDRADILYNI), 5'-CGGATCCGA(TC)(AC)GIGC(GATC)GA(TC)AT(ATC)(TC)T(GATC)TA(TC)AA(TC)AT-3' and used for PCR in combination with a λZAPII vector primer. PCR: 45 cycles of 94 °C, 20 s; 53 °C, 30 s; and 72 °C, 1 min. Hybridization for λZAPII CNS cDNA library screening was performed in 6× SSC (1× SSC: 0.15 M NaCl and 0.015 M Na-citrate), 0.2% SDS, 5× Denhardt's and 10 µg ml⁻¹ herring sperm DNA at 65 °C for 18 h. Washing was in 0.2× SSC, 0.2% SDS at 65 °C for 30 min.

Gel filtration

To determine the relative molecular mass of the multimer, recAChBP (45 µg) dissolved in running buffer (20 mM Tris/HCl, pH 8.0, 150 mM NaCl, 0.02% Na-azide) was run on a Superdex 200 HiLoad 16/60 column (Pharmacia). Marker proteins were run separately in running buffer (50 mM Tris/HCl, pH 8.0, 150 mM NaCl, 3 mM β-mercaptoethanol) on the same column. Absorbance detection was at 254 nm.

In situ hybridization and cytochemistry

For *in situ* hybridization, digoxigenin-labelled run-off sense and antisense cRNAs to the coding region of AChBP were synthesized using T7 or T3 RNA polymerase and digoxigenin-UTP labelling (Boehringer). Signal was detected by using sheep anti-digoxigenin, conjugated to alkaline phosphatase. For immunolabelling, a mouse polyclonal antibody raised against recAChBP was used, and detected by a sheep anti-mouse antibody (Dako), conjugated to horseradish peroxidase⁴⁹. For α-Bgt labelling sections were incubated in 1:100 diluted rhodamine-α-Bgt (Molecular Probes) in PBS for 60 min, washed 3 × 5 min in PBS and viewed with a fluorescence microscope.

Immunogold labelling

All studies were performed on adult specimens of *Lymnaea*. Ultrathin cryosections were prepared and (double)-immunolabelled by the protein A gold technique⁵⁰. Primary antibodies were a mouse polyclonal anti-AChBP and an affinity-purified polyclonal antibody against the α-caudodorsal cell peptide (anti-ELH). Grids were analysed in a Philips EM3000 electron microscope.

Sequence and mass determination

We homogenized 80 *Lymnaea* CNSs in lysis buffer (PBS (16 mM Na₂HPO₄, 4 mM NaH₂PO₄, pH 7.4, 150 mM NaCl), 0.5% Nonidet P-40, 0.1% Triton, 0.2% Tween-20) containing protease inhibitors. No detergents were used in making a soluble fraction. Lysates were cleared by centrifugation at 12,000g for 5 min. Streptavidin-coated magnetic beads (Dyna; 5 mg) were saturated with α-Bgt-Biotin (Molecular Probes), washed in PBS, incubated with the cleared lysate for 1 h and washed in PBS. In control α-Bgt was omitted. Proteins bound to α-Bgt-beads were eluted in 10 µl PBS, 10⁻⁴ M nicotine for 1 h. The eluent was separated on a microcolumn LC system and electrospray mass spectra were acquired on a mass spectrometer (Micromass, Q-TOF). Sequence analysis of α-Bgt binding protein was by the same protocol, followed by SDS–PAGE, western blotting on PVDF and Edman degradation using a protein sequencer (Model 473A, ABI).

Immunoprecipitation

Five *Lymnaea* CNSs were desheathed and mass dissociated in a culture dish, allowing glial cells to adhere selectively. Glia were incubated in 5 ml HEPES buffered saline (HBS; 5 mM

HEPES (pH 7.9), 51 mM NaCl, 1.7 mM KCl, 4.1 mM CaCl₂, 1.5 mM MgCl₂) containing 20 µCi ml⁻¹ [³⁵S]-Met/Cys mixture (Tran-S, ICN) at room temperature for 4 h and rinsed with HBS. α-Bgt was used at 10⁻⁷ M for 10 min and ACh at 10⁻⁵ M for 5 min, all in HBS. AChBP antibody was added to the releasate, incubated at 16 °C for 2 h, and Protein-A agarose (Gibco) was added for 1 h. Beads were rinsed with HBS, resuspended in SDS sample buffer and boiled, and the precipitated radiolabelled AChBP was analysed in triplicate on 8% SDS–PAGE and quantified using densitometry.

Received 4 January; accepted 19 April 2001.

- Araque, A., Parpura, V., Sanzgiri, R. P. & Haydon, P. G. Tripartite synapses: glia the unacknowledged partner. *Trends Neurosci.* **22**, 208–215 (1999).
- Smith, S. J. Do astrocytes process neural information? *Prog. Brain Res.* **94**, 119–136 (1992).
- Dani, J. W., Chernjavsky, A. & Smith, S. J. Neuronal activity triggers calcium waves in hippocampal astrocyte networks. *Neuron* **8**, 429–440 (1992).
- Pasti, L., Pozzan, T. & Carmignoto, G. Long-lasting changes of calcium oscillations in astrocytes. A new form of glutamate-mediated plasticity. *J. Biol. Chem.* **270**, 15203–15210 (1995).
- Parpura, V. et al. Glutamate-mediated astrocyte–neuron signaling. *Nature* **369**, 744–747 (1994).
- Nedergaard, M. Direct signaling from astrocytes to neurons in cultures of mammalian brain cells. *Science* **263**, 1768–1771 (1994).
- Murphy, T. H., Blatter, L. A., Wier, W. G. & Baraban, J. M. Rapid communication between neurons and astrocytes in primary cortical cultures. *J. Neurosci.* **13**, 2672–2679 (1993).
- Newman, E. A. & Zahs, K. R. Modulation of neuronal activity by glial cells in the retina. *J. Neurosci.* **18**, 4022–4028 (1998).
- Pasti, L., Volterra, A., Pozzan, T. & Carmignoto, G. Intracellular calcium oscillations in astrocytes: a highly plastic, bidirectional form of communication between neurons and astrocytes in situ. *J. Neurosci.* **17**, 7817–7830 (1997).
- Cornell-Bell, A. H., Finkbeiner, S. M., Cooper, M. S. & Smith, S. J. Glutamate induces calcium waves in cultured astrocytes: long-range glial signaling. *Science* **247**, 470–473 (1990).
- Calegari, F. et al. A regulated secretory pathway in cultured hippocampal astrocytes. *J. Biol. Chem.* **274**, 22539–22547 (1999).
- Porter, J. T. & McCarthy, K. D. Hippocampal astrocytes in situ respond to glutamate released from synaptic terminals. *J. Neurosci.* **16**, 5073–5081 (1996).
- Araque, A., Parpura, V., Sanzgiri, R. P. & Haydon, P. G. Glutamate-dependent astrocyte modulation of synaptic transmission between cultured hippocampal neurons. *Eur. J. Neurosci.* **10**, 2129–2141 (1998).
- Araque, A., Sanzgiri, R. P., Parpura, V. & Haydon, P. G. Calcium elevation in astrocytes causes an NMDA receptor-dependent increase in the frequency of miniature synaptic currents in cultured hippocampal neurons. *J. Neurosci.* **18**, 6822–6829 (1998).
- Kang, J., Jiang, L., Goldman, S. A. & Nedergaard, M. Astrocyte-mediated potentiation of inhibitory synaptic transmission. *Nature Neurosci.* **1**, 683–692 (1998).
- Jahromi, B. S., Robitaille, R. & Charlton, M. P. Transmitter release increases intracellular calcium in perisynaptic Schwann cells in situ. *Neuron* **8**, 1069–1077 (1992).
- Robitaille, R., Jahromi, B. S. & Charlton, M. P. Muscarinic Ca responses resistant to muscarinic agonists at the perisynaptic Schwann cells at the frog neuromuscular junction. *J. Physiol.* **504**, 337–347 (1997).
- Robitaille, R. Modulation of synaptic efficacy and synaptic depression by glial cells at the frog neuromuscular junction. *Neuron* **21**, 847–855 (1998).
- Theodosis, D. T. & Poullain, D. A. Activity-dependent neuronal glial and synaptic plasticity in the adult mammalian hypothalamus. *Neurosci.* **57**, 501–535 (1993).
- Brejck, K. et al. Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. *Nature* **411**, 269–276 (2001).
- Feng, Z. P., Klumperman, J., Lukowiak, K. & Syed, N. I. In vitro synaptogenesis between the somata of identified *Lymnaea* neurons requires protein synthesis but not extrinsic growth factors or substrate adhesion molecules. *J. Neurosci.* **17**, 7839–7849 (1997).
- Hamakawa, T. et al. Excitatory synaptogenesis between identified *Lymnaea* neurons requires extrinsic trophic factors and is mediated by receptor tyrosine kinases. *J. Neurosci.* **19**, 9306–9312 (1999).
- Woodin, M. A., Hamakawa, T., Takasaki, K. & Syed, N. I. Trophic factor-induced plasticity of synaptic connections between identified *Lymnaea* neurons. *Learn. Mem.* **6**, 307–316 (1999).
- Orr-Urteger, A. et al. Mice deficient in the alpha7 neuronal nicotinic acetylcholine receptor lack alpha bungarotoxin binding sites and hippocampal fast nicotinic currents. *J. Neurosci.* **17**, 9165–9171 (1997).
- Gardner, D. & Kandel, E. R. Interconnections of identified multiaction interneurons in buccal ganglia of aplysia. *J. Neurophys.* **40**, 349–361 (1977).
- Yeoman, M. S., Parish, D. C. & Benjamin, P. R. A cholinergic modulatory interneuron in the feeding system of the snail, *Lymnaea*. *J. Neurosci.* **70**, 37–50 (1993).
- Le Novere, N. & Changeux, J. P. Molecular evolution of the nicotinic acetylcholine receptor: an example of multigene family in excitable cells. *J. Mol. Evol.* **40**, 155–172 (1995).
- Devillers-Thiery, A., Galzi, J. L., Eiselé, J. L., Bertrand, S. & Changeux, J. P. Functional architecture of the nicotinic acetylcholine receptor: a prototype of ligand-gated ion channels. *J. Membrane Biol.* **136**, 97–112 (1993).
- Karlin, A. & Akabas, M. H. Toward a structural basis for the function of nicotinic acetylcholine receptors and their cousins. *Neuron* **6**, 1231–1244 (1995).
- Mehta, A. K. & Ticku, M. K. An update on GABA_A receptors. *Brain Res. Rev.* **2–3**, 196–217 (1999).
- Betz, H. et al. Structure and functions of inhibitory and excitatory glycine receptors. *Annu. NY Acad. Sci.* **868**, 667–676 (1999).
- Corringer, P. J., Le Novere, N. & Changeux, J. P. Nicotinic receptors at the amino acid level. *Annu. Rev. Pharmacol. Toxicol.* **40**, 431–458 (2000).
- Drisdell, R. C. & Green, W. N. Neuronal α-Bungarotoxin receptors are α7 subunit homomers. *J. Neurosci.* **20**, 133–139 (2000).
- Green, W. N. Channel assembly: creating structures that function. *J. Gen. Physiol.* **113**, 163–169 (1999).
- Verall, S. & Hall, Z. W. The N-terminal domain of the acetylcholine receptor subunits contain recognition signals for the initial steps of receptor assembly. *Cell* **68**, 23–31 (1992).
- Wang, Z. Z., Hardy, S. F. & Hall, Z. W. Assembly of the nicotinic acetylcholine receptor. The first

- transmembrane domains of truncated alpha and delta subunits are required for heterodimer formation *in vivo*. *J. Biol. Chem.* **271**, 27575–27584 (1996).
37. Wells, G. B., Anand, R., Wang, F. & Lindstrom, J. Water-soluble nicotinic acetylcholine receptor formed by alpha7 subunit extracellular domains. *J. Biol. Chem.* **273**, 964–973 (1998).
38. Corringer, P. J. *et al.* Critical elements determining diversity in the agonist binding and desensitization of neuronal nicotinic acetylcholine receptors. *J. Neurosci.* **18**, 648–657 (1998).
39. Badio, B. & Daly, J. W. Epibatidine, a potent analgetic and nicotinic agonist. *Mol. Pharmacol.* **45**, 563–569 (1994).
40. Groebe, D. R. & Abramson, S. N. Lophotoxin is a slow binding irreversible inhibitor of nicotinic acetylcholine receptors. *J. Biol. Chem.* **270**, 281–286 (1995).
41. Edelstein, S. J. & Bardsley, W. G. Contributions of individual molecular species to the Hill coefficient for ligand binding by an oligomeric protein. *J. Mol. Biol.* **267**, 10–16 (1997).
42. Zwart, R. & Vijverberg, H. P. Potentiation and inhibition of neuronal nicotinic receptors by atropine: competitive and noncompetitive effects. *Mol. Pharmacol.* **52**, 886–895 (1997).
43. Verbitsky, M., Rothlin, C. V., Katz, E. & Belen Elgoyhen, A. Mixed nicotinic–muscarinic properties of the alpha9 nicotinic cholinergic receptor. *Neuropharmacology* **39**, 2515–2524 (2000).
44. Whatey, J. C., Nass, M. M. & Lester, H. A. Numerical reconstruction of the quantal event at nicotinic synapses. *Biophys. J.* **27**, 145–164 (1979).
45. Land, B. R., Salpeter, E. E. & Salpeter, M. M. Kinetic parameters for acetylcholine interaction in intact neuromuscular junction. *Proc. Natl Acad. Sci. USA* **78**, 7200–7204 (1981).
46. Katz, B. & Miledi, R. The binding of acetylcholine to receptors and its removal from the synaptic cleft. *J. Physiol.* **231**, 549–574 (1973).
47. Kuffler, S. W. & Yoshikama, D. The number of transmitter molecules in a quantum: an estimate from iontophoretic application of acetylcholine at the neuromuscular synapse. *J. Physiol.* **1**, 465–482 (1975).
48. Syed, N. I., Bulloch, A. G. & Lukowiak, K. *In vitro* reconstruction of the respiratory central pattern generator of the mollusk *Lymnaea*. *Science* **250**, 282–285 (1990).
49. Van Minnen, J., Van De Haar, C., Raap, A. K. & Vreugdenhil, E. Localization of ovulation hormone-like neuropeptide in the central nervous system of the snail *Lymnaea stagnalis* by means of immunocytochemistry and *in situ* hybridization. *Cell Tissue Res.* **251**, 477–484 (1988).
50. Slot, J. W., Geuze, H. J. & Weerkamp, A. J. Localization of macromolecular components by application of the immunogold technique on cryosectioned bacteria. *Methods Microbiol.* **20**, 211–236 (1988).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank S. N. Abramson for bipinnatin-B; E. R. Van Kesteren and Broers-Vendrig for immunocytochemistry; and F. v. Huizen for suggestions and reagents for the pharmacological experiments. A.B.S. and K.B. were supported by a grant of the Organization for Applied Research (STW). N.I.S. is an Alberta Heritage Foundation for Medical Research senior scholar and was supported by CIHR and NSERC (Canada), and by the Dutch NWO.

Correspondence and request for materials should be addressed to A.B.S. (e-mail: absmit@bio.vu.nl). AChBP Genbank accession number: AF364899.

Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors

Katjuša Brejc*, Willem J. van Dijk*, Remco V. Klaassen†, Mascha Schuurmans†, John van der Oost†‡, August B. Smit† & Titia K. Sixma*

* Division of Molecular Carcinogenesis, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

† Department of Molecular and Cellular Neurobiology, Research Institute Neurosciences Vrije Universiteit, Faculty of Biology, De Boelelaan 1087, 1081 HV Amsterdam, The Netherlands

Pentameric ligand gated ion-channels, or Cys-loop receptors, mediate rapid chemical transmission of signals. This superfamily of allosteric transmembrane proteins includes the nicotinic acetylcholine (nAChR), serotonin 5-HT₃, γ -aminobutyric-acid (GABA_A and GABA_C) and glycine receptors. Biochemical and electrophysiological information on the prototypic nAChRs is abundant but structural data at atomic resolution have been missing. Here we present the crystal structure of molluscan acetylcholine-binding protein (AChBP), a structural and functional homologue of the amino-terminal ligand-binding domain of an nAChR α -subunit. In the AChBP homopentamer, the protomers have an immunoglobulin-like topology. Ligand-binding sites are located at each of five subunit interfaces and contain residues contributed by biochemically determined 'loops' A to F. The subunit interfaces are highly variable within the ion-channel family, whereas the conserved residues stabilize the protomer fold. This AChBP structure is relevant for the development of drugs against, for example, Alzheimer's disease and nicotine addiction.

The superfamily of pentameric ligand-gated ion channels (LGICs), including nAChR, 5-HT₃, GABA_A and GABA_C, and glycine receptors, mediates chemical synaptic transmission. All of these LGICs form homo- or heteropentamers of related subunits, implying a common evolutionary origin¹. Each subunit is divided into an N-terminal or ligand-binding domain (LBD), a transmembrane region and an intracellular region². The LBDs, which are around 210 amino-acid residues long, contain ligand-binding sites for agonists and competitive antagonists. Binding of an agonist induces rapid opening of the transmembrane ion channel, leading to a change in membrane potential. The signal transmission mechanism is explained by allosteric interactions between subunits, involving numerous conformational transitions^{3,4}.

The nAChRs are extensively studied and can be divided into muscle and neuronal types². The muscle type, with stoichiometry (α_1)₂(β_1)₂(γ)₂, is found at the neuromuscular junction and in the electric organs of fish such as the electric ray *Torpedo californica*. It presents a common drug target for muscle relaxants. The neuronal nAChRs, which are hetero- or homomeric (for example, (α_4)₂(β_2)₃ or (α_7)₅), are located on both pre- and postsynaptic nerve terminals. They are important drug targets as they mediate nicotine addiction in smokers and the positive effects of nicotine on cognition, memory and attention in patients with, for example, schizophrenia or Alzheimer's and Parkinson's diseases⁵.

In nAChRs, the ligand-binding site is located at the interface between two subunits^{2,4}. Numerous biochemical studies showed that the principal part is always formed by the α -subunit residues—contributing to the so-called 'loops' A (ref. 6), B (ref. 7) and C (refs 7–10)—whereas the neighbouring subunit residues—contributing to the 'loops' D (refs 11, 12), E (refs 10, 13) and F (refs 14–16)—form the complementary part of the binding pocket. Thus, the muscle nAChR contains two different ligand-binding sites ($\alpha_1\delta$ and $\alpha_1\gamma$) with distinct affinities, whereas the homopentameric α_7 receptor contains five identical ligand-binding sites. In these sites

acetylcholine is expected to bind through cation- π interactions, where the positive charge of the quaternary ammonium of acetylcholine interacts with the electron-rich aromatic side chains¹⁷.

So far, the only experimental structural information on the nAChRs came from electron microscopy studies on *Torpedo* receptors^{18–20}. These low–medium-resolution studies revealed the overall shape and dimensions of the receptor, a location of the ligand-binding site and the organization of the ion channel. The most recent 4.6 Å data²¹ showed acetylcholine-binding pockets surrounded by seven-stranded β -sheet structures. This fits with circular dichroism measurements demonstrating that LBDs are predominantly β -structured^{22–24}. No high-resolution data are available yet and although soluble LBDs have been produced^{22–26}, protein quantities were insufficient for high-resolution studies.

Acetylcholine-binding protein

Acetylcholine-binding protein (AChBP) is a soluble protein found in the snail *Lymnaea stagnalis*²⁷. It is produced and stored in glial cells, and released in an acetylcholine-dependent manner into the synaptic cleft, where it modulates synaptic transmission. Mature AChBP is 210 residues long and forms a stable homopentamer. It aligns with the N-terminal domains of pentameric LGICs (Fig. 1) and lacks the transmembrane and intracellular domains present in the superfamily²⁷. AChBP is most closely related to the α -subunits of the nAChRs (Fig. 1). Nearly all residues that are conserved within the nAChR family are present in AChBP, including those that are relevant for ligand binding. Moreover, AChBP binds known nAChR agonists and competitive antagonists such as acetylcholine, nicotine, *d*-tubocurarine and α -bungarotoxin²⁷. Therefore, AChBP can be used as an example of the N-terminal domain of an α -subunit of nAChRs. Here we report the crystal structure of the AChBP homopentamer at 2.7 Å resolution.

Structure determination

The crystal structure of AChBP was solved using weak Pb multiple-wavelength anomalous diffraction (MAD) data in two crystal forms.

‡ Present address: Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University, Hesselink van Suchtelenweg 4, 6703 CT Wageningen, The Netherlands.

The electron density map was improved substantially by cross-crystal averaging of three crystal forms with 20, 10 and 5 copies of the protomer in the asymmetric unit (Table 1). The structure was refined at 2.7 Å in space group $P4_21_2$, with one AChBP pentamer in the asymmetric unit. Refinement with partial five-fold non-crystallographic symmetry (NCS) restraints resulted in an R -factor of 26.4% ($R_{\text{free}} = 30\%$).

The AChBP pentamer

The AChBP homopentamer, when viewed along the five-fold axis, resembles a windmill toy, with petal-like protomers (Fig. 2a). When viewed perpendicular to the five-fold axis it forms a 62-Å-high cylinder (Fig. 2b), with a diameter of 80 Å. The diameter of the

central hole is ~ 18 Å, between side chains. These dimensions are in good agreement with the N-terminal domain in the *Torpedo* nAChR electron microscopy data²¹. In the pentamer the only subunit contacts are dimer interfaces, of which each protomer has two, the plus side and the minus side. We refer to the A (plus)–B (minus) interface as an example for the five equivalent interfaces AB, BC, CD, DE and EA (Fig. 2).

The AChBP protomer

Each AChBP protomer is a single domain protein, asymmetric in shape, with a size of around $62 \times 47 \times 34$ Å³ (Fig. 3a). It consists of an N-terminal α -helix, two short 3_{10} helices and a core of ten β -strands, which form a β -sandwich. The order of β -strands conforms

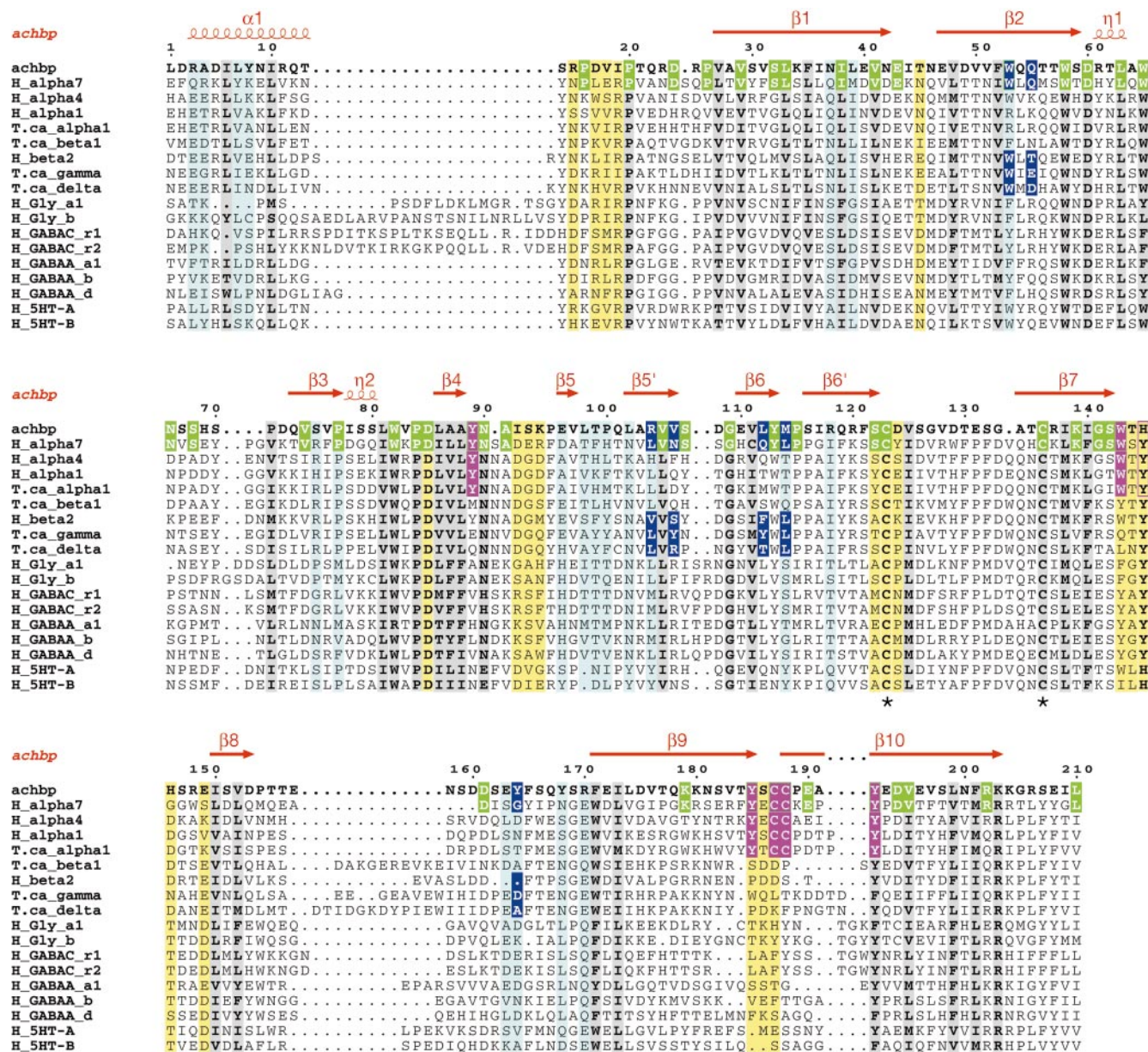


Figure 1 Sequence alignment of AChBP with pentameric ligand-gated ion channels (LGICs). The alignment shows only the N-terminal domain of the LGIC subunits and is based on a multi-sequence ClustalX⁵⁰ alignment of 92 full-length pentameric LGIC sequences. Alignments are shown from first to last AChBP residue (for example, starting at second residue of mature *Torpedo* α , β γ and δ subunits, ending ~ 4 residues into the first transmembrane domain²⁹). H: human, T.ca: *Torpedo californica*. Secondary structure elements (α : α -helix, β : β -strand, η : 3_{10} -helix) are indicated in red above the sequence, in accordance with Fig. 3a. Precise beginnings and ends will change with

higher resolution. AChBP shares 24% sequence identity with the ligand-binding domain (LBD) of human α_7 (shown in green), 20–24% with other nicotinic receptors and 15–18% with other LGICs. The residues conserved in the superfamily are shown in bold with grey background. Asterisk, beginning and end of the Cys loop. Colouring of interface residues at the plus (yellow) and minus side (light blue) shows the lack of sequence conservation in the subunit interface across the pentameric LGIC family. Nicotinic receptor ligand-binding residues on the principal (pink) and complementary (dark blue) side are indicated.

Table 1 Data collection statistics

Data set	λ_1 peak	λ_2 remote	λ_3 infl.	λ_1 peak	λ_2 infl.	Native
Space group	$P2_12_12$			$P2_1$		$P4_22_12$
Resolution (Å)	3.3/3.4–3.3			3.0/3.1–3.0		2.7/2.8–2.7
Wavelength (Å)	0.9492	0.8610	0.9507	0.9479	0.9498	0.943
Completeness (%)	99.7/99.7	99.6/99.6	99.7/99.7	99.9/99.9	99.5/99.5	97.8/96.5
Mosaicity (deg.)	0.62			0.43		0.78
Redundancy	3.7/3.8	3.8/3.9	3.7/3.8	3.5/2.2	3.2/2.0	6.5
R_{merge} (%)	7.7/46.8	7.8/45.2	8.3/55.0	5.9/26.1	6.0/32.9	5.9/67.4
$I/\sigma I$	8.7/1.6	8.4/1.7	8.3/1.4	7.7/2.7	6.8/1.5	27.4/2.3
Phasing	ISO/ANO	ISO/ANO	ISO/ANO	ISO/ANO	ISO/ANO	
R_{cullis} (%)	0.74/0.89	n.a./0.92	0.54/0.94	n.a./0.74	0.66/0.77	
Phasing power	0.57/1.2	n.a./1.06	2.3/0.91	n.a./1.1	0.37/1.22	
Figure of merit (overall)	0.45			0.28		

$\langle I/\sigma I \rangle$, Mean signal to noise, where I is the integrated intensity of a measured reflection and σI is the estimated error in the measurement; $R_{\text{merge}} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$; $R_{\text{cullis}} = \sum_{hkl} |F_{\text{PH}} + F_P| - F_{\text{H}}(\text{calc}) / \sum_{hkl} |F_{\text{PH}} + F_P|$.

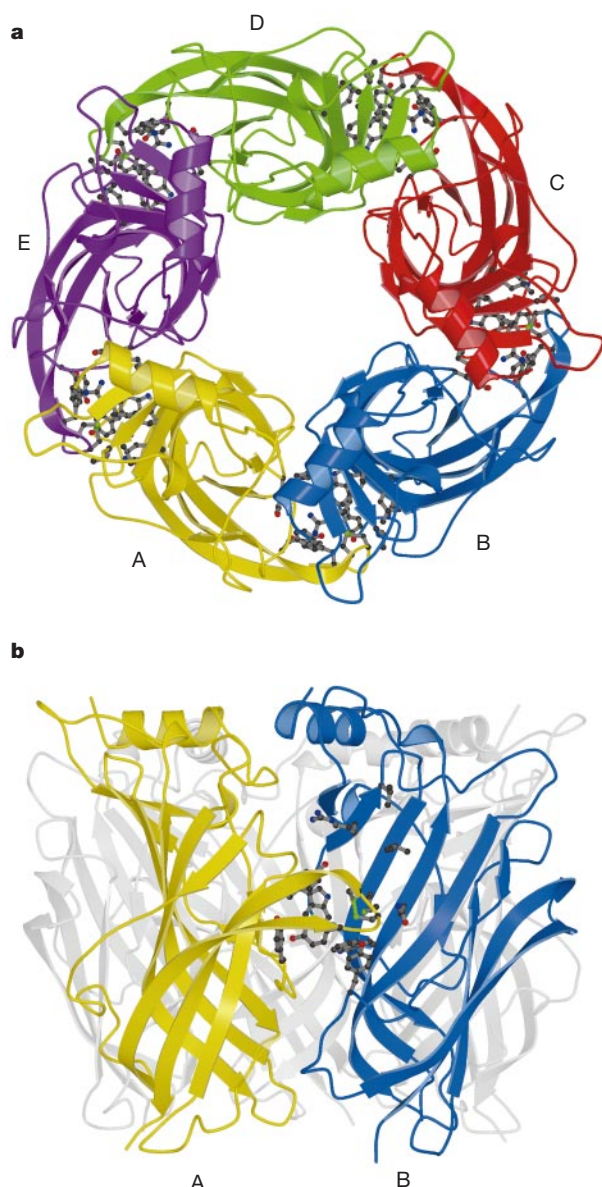


Figure 2 The pentameric structure of AChBP. **a**, In this representation each protomer has a different colour. Subunits are labelled anti-clockwise, with A–B, B–C, C–D, D–E and E–A forming the plus and minus interface side, with the principal and complementary ligand-binding sites, respectively (ball-and-stick representation). **b**, Viewing the AChBP pentamer perpendicular to the five-fold axis. The equatorially located ligand-binding site (ball-and-stick representation) is highlighted only in the A (yellow)–B (blue) interface.

to a modified immunoglobulin (Ig) topology²⁸ (Fig. 3b) with an extra β -hairpin (f' – f'') and an extra strand (b'). These additional strands introduce two ‘Greek key’ folding motifs. An Ig-like topology had been predicted for nAChRs^{2,29}, but the location of the ligand-binding site was missed, owing to the presence of extra β -strands. Compared with the classical Ig-fold²⁸, the AChBP β -strands are considerably twisted, with the β -sheets rotated against each other, resulting in two separate hydrophobic cores. Thus the three-dimensional fold does not resemble other Ig-like proteins and comparison³⁰ with the protein database did not result in a significant match to any known structure.

Positioning of functional regions

In the structure, the N and C termini are located at ‘top’ and ‘bottom’ of the pentamer, respectively. In the ion channels the transmembrane domains are at the C-terminal end of the LBD, at the ‘bottom’ of the AChBP structure (Figs 2b and 3), starting directly at the end of β -strand $\beta 10$.

In muscle type nAChRs, the main immunogenic region (MIR), comprising residues $\alpha_1 67$ – $\alpha_1 76$, acts as an epitope in the autoimmune disease myasthenia gravis³¹. Although the MIR-related region in AChBP (residues 65–72) has no sequence homology with the α_1 -subunit, its location in a highly accessible position in loop L3 at the ‘top’ of the pentamer agrees with the expected accessibility for this region (Fig. 3a). It also fits with electron microscopy studies that located the MIR at the distal end of the receptor relative to the membrane²⁰.

The central pore of the pentamer is very hydrophilic, lined with charged residues. On each AChBP promoter, a large pocket is visible, accessible from the central pore. Each pocket, framed at the entrance by β -strands ($\beta 3$, $\beta 4$, $\beta 5$ and $\beta 5'$) (Fig. 3a), is uncharged and mainly hydrophobic. This region probably corresponds to the tunnel framed by twisted β -strands that was observed in the α_1 -subunit of the *Torpedo* receptor at 4.6 Å resolution²¹.

There is a different cavity at each interface between the subunits. These cavities are lined by residues, which were biochemically shown to be involved in ligand binding in nAChRs^{2,5–16}. These cavities are mostly buried from the solvent, and located close to the outside of the pentameric ring (Fig. 2a). When viewed perpendicular to the five-fold axis, they are roughly equatorially positioned, about 30 Å away from the C termini (Fig. 2b), conforming to the expected location of the *Torpedo* receptor ligand-binding site, as determined by labelling³² and electron microscopy studies¹⁸. We have concluded that these cavities are the ligand-binding sites.

The ligand-binding site

Each ligand-binding site is found in a cleft formed by a series of loops from the principal face of one subunit and a series of β -strands from the complementary face of an adjacent subunit (Fig. 4). The

principal side on the plus side of the AB interface consists of residues coming from loop A (Tyr A89), loop B (Trp A143, A145) and loop C (Tyr A185, the double cysteine A187–A188 and Tyr A192) (Fig. 4c). The complementary part of this binding site is formed by β -strands in protomer B contributing loop D (Trp B53, Gln B55), loop E (Arg B104, Val B106, Leu B112 and Met B114) and loop F (Tyr B164) (Fig. 4d). Four of the aromatic residues form the bottom half of the cavity (Tyr A89, Tyr A185, Tyr B164 and Trp B53). The walls are formed by Tyr A192, Trp A143, the main chain of A145, the side chains of Met B114 and Gln B55, and the vicinal disulphide (Cys A187–Cys A188). The hydrophobic parts of Arg B104, Val B106 and Leu B112 form the top of the binding site (Fig. 4a).

All residues in the binding site have been identified by photo-affinity labelling and mutagenesis studies^{6–16}. Although the side chain of His A145 is pointing away from the cavity, its main chain is involved in the binding site. One residue identified by labelling studies⁶, Trp A82 (α_1 Trp 86), is involved in hydrophobic core formation and located far from the pocket. Therefore it probably does not participate directly in ligand binding. Additional residues may be involved in binding large ligands such as toxins¹³. Otherwise, the AChBP ligand-binding site entirely confirms the available biochemical and mutational data on nAChRs. The structure, however, reveals for the first time how these residues are positioned with respect to each other, and will provide a valuable tool for drug design studies.

All observed residues are conserved between known ligand-binding subunits of nicotinic receptors except the loop F Tyr B164 residue. In our structure the loop F region has an unusual conformation, but as it is relatively weakly resolved, its precise analysis is difficult. The loop F region has low sequence conservation in the nicotinic family (Fig. 1), and in other superfamily members it may have a different conformation, providing different residues to the binding site. Such changes could lead to variations in affinity, for example by changing the size of the ligand-binding site or its access route.

Close to the loop F region, the electron density was interpreted as a calcium ion, because the crystals were grown in the presence of

Ca^{2+} . The cation has Asp B161, Asp B175 and the main chain of B176 as ligands. This putative Ca^{2+} ion may help to orientate Tyr B164 and could thus be involved in ligand binding. This agrees with observations that the residue equivalent to Asp B161 in muscle/*Torpedo* subunits (γ Asp 174/ δ Asp 180) is important in ligand binding^{14,15}. Additionally, calcium-binding sites that enhance the response to agonist binding have been identified in a homologous region (residue range 161–172) of the neuronal α_7 -receptor³³.

The most likely access routes to the ligand-binding sites are from above or below the double-cysteine-containing loop C (Fig. 4a). This region buries the ligand-binding site from the solvent, preventing access from the outside. Access of some of the larger ligands, such as *d*-tubocurarine, would require the binding site to be opened up, for example by movement of the β -hairpin β_9 – β_{10} in the C-loop region. Access from the central pore has been suggested²¹, but this would require major structural rearrangements at the interface, which are less likely.

From the location of the ligand-binding site, we can draw conclusions about the arrangement of ligand-binding α_1 -subunits with respect to their complementary partners in the *Torpedo* and muscle receptors. It has been suggested that the $\alpha_1\gamma$ and $\alpha_1\delta$ interfaces occur in a clockwise $\alpha_1\gamma\alpha_1\delta\beta_1$ arrangement when looking towards the membrane³⁴. Our structure would favour an anticlockwise $\alpha_1\gamma\alpha_1\delta\beta_1$ arrangement, as the principal ligand-binding site is located on the anticlockwise side of its complementary partner (Fig. 2).

Ligand binding

To our surprise we found features of bulky electron density that stacked onto Trp 143 in each ligand-binding site in the experimental cross-crystal averaged electron density (Fig. 4b). We have assigned this to a HEPES (*N*-2-hydroxyethylpiperazine-*N'*-2-ethanesulphonic acid) buffer molecule, which contains a positively charged quaternary ammonium group and therefore has some similarity to known nicotinic receptor ligands. Its dissociation constant of 100 mM (data not shown) indicates that binding under crystallization conditions (100–150 mM) is possible at low occupancy.

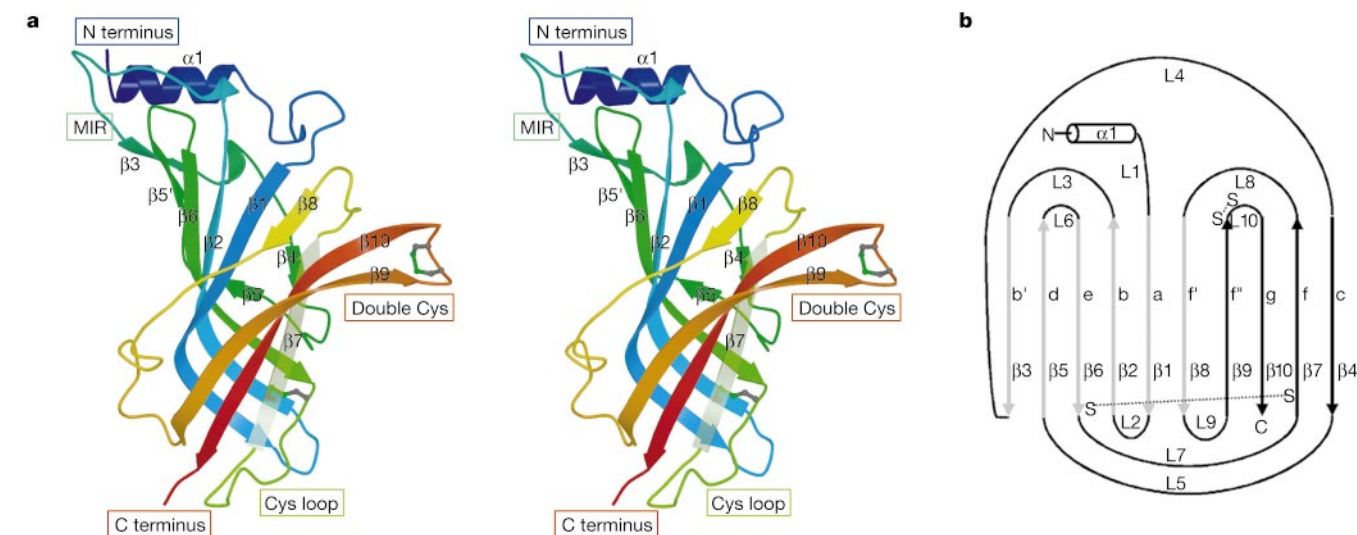


Figure 3 Overview of the AChBP protomer structure. **a**, Stereo representation of the AChBP protomer as viewed from outside the pentameric ring. This ribbon representation is coloured as a rainbow gradient, from blue (N terminus) to red (C terminus). Disulphide bridges are indicated in green ball-and-stick representation. In a complete ion channel the N terminus would point towards the synaptic cleft and the C terminus would enter the membrane at the bottom, continuing into the first transmembrane domain. **b**, Topology diagram of the AChBP protomer. For comparison with Ig-folds the strands have been

labelled a–g, showing the additional strand (b') and hairpin (f'–f''). In this structure, strands have been labelled β_1 – β_{10} with loops (or turns) L1–L10 preceding each strand with the same number. The β_5 strand is broken (β_5 – β_5') with internal loop L5'; β_6 also has a small break, but it is shown continuously (see Fig. 1). The precise beginnings and ends of strands may change slightly with increasing resolution, but the topology seen here will be highly conserved across the entire family of pentameric LGICs. S: disulphide bridge.

Although HEPES does not make any specific hydrogen bonds with the protein, it stacks with its quaternary ammonium onto Trp 143, making cation- π interactions as expected for nicotinic agonists^{17,35} (Fig. 4b). It buries equal amounts of surface of the principal and complementary subunit. However, owing to low occupancy and limited resolution of the data, the precise orientation of the HEPES molecule cannot be definitely resolved. It has been suggested³ that the ligand-binding site of nAChRs could be similar to that of acetylcholinesterase (AChE). Although the size of the binding site is roughly similar in AChBP and AChE, the observed arrangement of aromatic residues is quite different. However, the stacking of the quaternary ammonium of HEPES, as far as it can be refined in the current AChBP structure, is similar to that of the quaternary ammonium of the decamethonium in AChE on Trp 84 (ref. 36).

Pentamer interface

The subunit interface consists on the plus side entirely of loop regions (L1, L2, L4, L5, L7, L8 and L10), whereas the minus side mostly presents secondary structure elements to the interface (α 1,

β 1, β 2, β 3, β 5, β 6 and L9; Fig. 5). Several residues are important for both ligand binding and pentamer formation. The interface buries a large surface area (2,700 Å²), with a mainly uncharged character including only a single bifurcated salt bridge (Glu A149–Arg B3 and Arg B104). It is a very convoluted surface, indicating that shape complementarity may be important in pentamer formation (Fig. 5a). The interface residues are not well conserved between subfamilies of the superfamily of pentameric ligand-gated ion channels (Fig. 1). The residues change markedly between the different ion-channel subtypes, with for example changes from hydrophobic to charged and vice versa (Fig. 1).

Ligand-gated ion channels

The superfamily of ligand-gated ion channels has highly conserved LBDs, and the function and location of the conserved residues can now be analysed in the light of the structure. Most residues that are conserved in the superfamily (Fig. 1) are hydrophobic and help to maintain the hydrophobic core of the protomer, grouped into three clusters (Fig. 6). The first cluster is involved in packing of

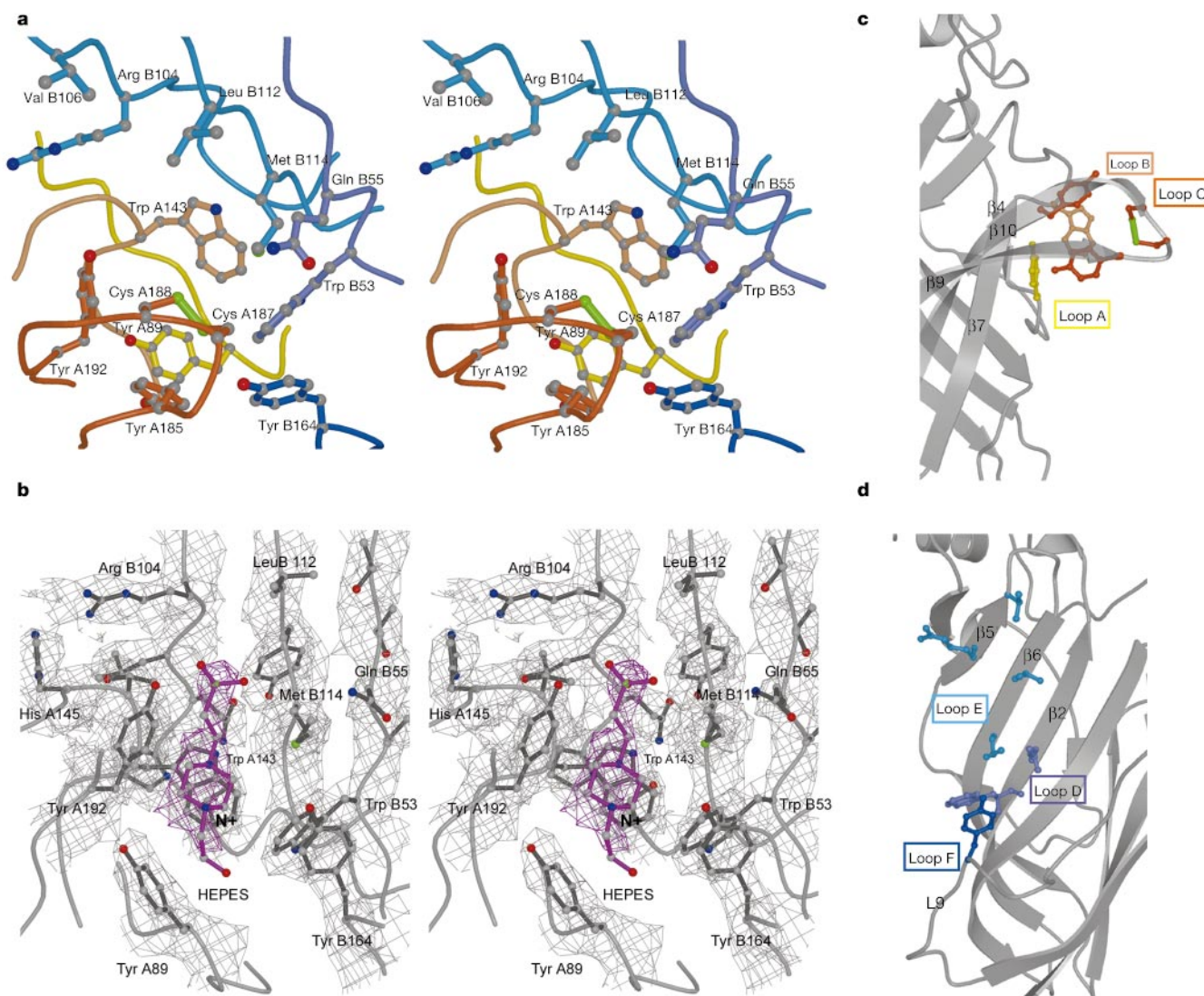


Figure 4 The ligand-binding site. **a**, Stereo representation of the ligand-binding site in ball-and-stick representation, showing the contribution of the principal 'loops': A (Tyr A89/ α 1 Tyr 93, yellow), B (Trp A143/ α 1 Trp 149, dark yellow) and C (Tyr A185/ α 1 Tyr 190, Cys A187/ α 1 Cys 192, Cys A188/ α 1 Cys 193, Tyr A192/ α 1 Tyr 198, orange), and the complementary 'loops' D (Trp B53/ γ Trp 55, Gln B55/ γ Glu 57, violet), E (Arg B104/ γ Leu 109, Val B106/ γ Tyr 111, Leu B112/ γ Tyr 117, Met B114/ γ Leu 119,

light blue) and F (Tyr B164, blue). **b**, Stereo view of the electron density map displaying a HEPES buffer molecule in the ligand-binding site. This experimental density (contoured at 1 σ) is derived from cross-crystal averaging. **c**, Location of the principal ligand-binding residues (colours as in **a**, orientation as in Fig. 2b). **d**, Location of the complementary ligand-binding residues (colours as in **a**, orientation as in Fig. 2b). Note that loops D, E and F are all on β -strands.

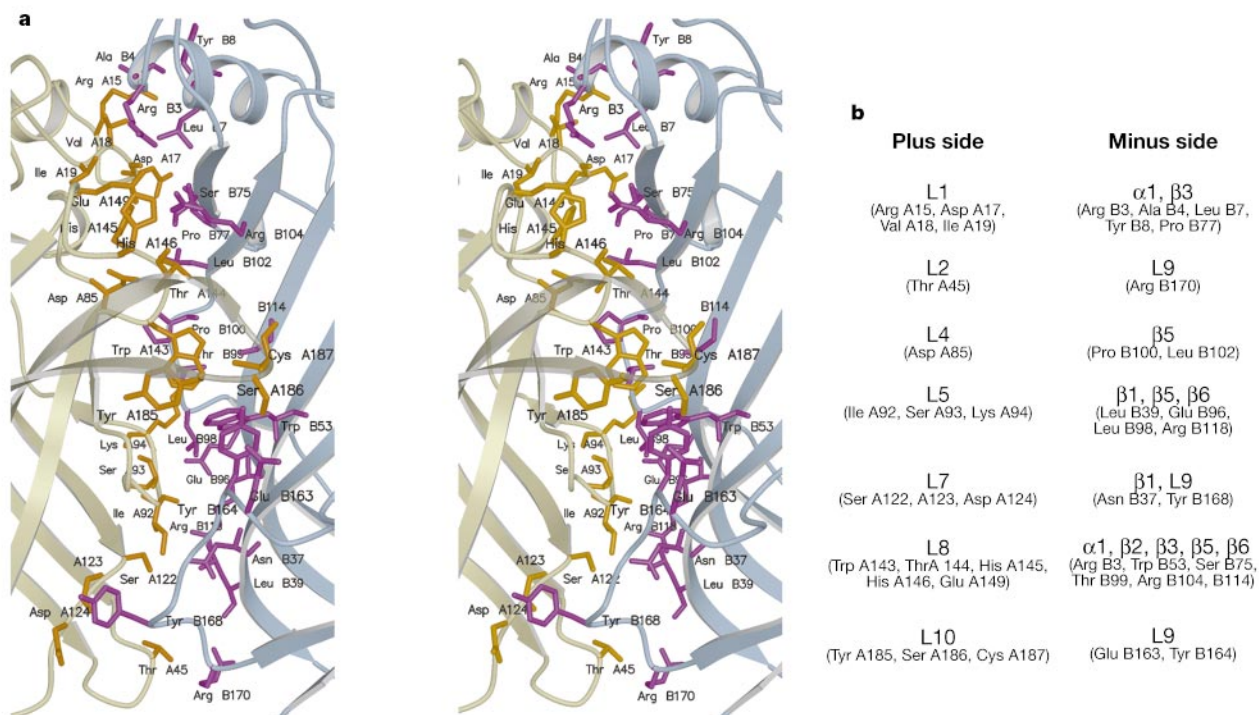


Figure 5 Dimer interface. **a**, Stereo figure of the dimer interface. Representation of the interface residues (ball-and-stick) on a schematic secondary structure. The figure shows the plus face of subunit A (light yellow, orange interface residues) and the minus face of subunit B (light blue, pink interface side chains). **b**, Dimer interface interactions. Note that

owing to the low conservation of these interfaces (Fig. 1) the actual interactions will not be conserved in any pentameric LGIC interface, but that in all receptors these topological regions are likely to form the interface.

the N-terminal helix $\alpha 1$ against the main framework of the protomer. The second cluster is situated in the upper half of the β -core region. The third cluster is located at the lower end of the β -sandwich (Fig. 6). In the superfamily, two residues are conserved in the ligand-binding site. Only very few conserved residues are hydrophilic; two of these maintain the turns of a Greek key motif connecting strands $\beta 3$, $\beta 5$, $\beta 6$ and $\beta 2$, in which Asp 60 stabilizes the N terminus of a small 3_{10} helix and Gly 109 enables tight-turn formation. Conserved residues Asp 85 and Asn 90 are involved in packing of the β -sheets. Asp 85 forms hydrogen bonds to the highly conserved Ser 142 and Thr 144, and residue Asn 90 brings together the main-chain oxygens of Ser 122 and Arg 137, enabling disulphide-bond formation of the nearby absolutely conserved disulphide bond (123–136). This disulphide bond is topologically equivalent to the ‘tyrosine cornerstone’³⁷ found in Ig-like proteins, in which tyrosine links the two β -sheets together, and a similar role is fulfilled by the disulphide bond in AChBP. This structural role for the disulphide bond explains why, in the *Torpedo* receptor, the Cys 128–Cys 142 bond is important for both preservation of subunit conformational stability³⁸ and complete nAChR assembly³⁹. As the conserved residues mainly contribute to the overall structure formation, it is clear that all pentameric LGIC N-terminal domains will have the same three-dimensional structure.

Contrary to the above residues, the Cys loop (Fig. 3a) is well conserved in the pentameric LGIC family but not in AChBP (Fig. 1, residues 129–141). It is hydrophobic in the receptors, whereas it is hydrophilic in AChBP. The Cys loop is located at the bottom (membrane) side of the protein, close to the dimer interface. This position, and its hydrophobicity in the LGIC family, implies that it could interact with the membrane or with the transmembrane region of the receptors, functions that are absent in AChBP.

Finally, the interface region is among the least conserved regions in the superfamily (Fig. 1). Pentamer formation does not apparently impose very stringent evolutionary requirements on these domains.

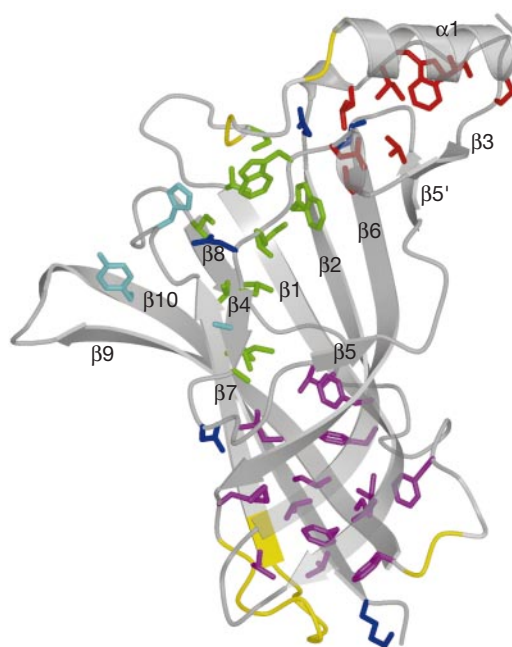


Figure 6 Conservation in the pentameric LGIC superfamily. Conserved residues are indicated as viewed from the central pore. Hydrophobic cluster I (red): residues 6, 10, 63, 65, 71, 81, 105, 111; Cluster II (green): residues 20, 27, 29, 58, 82, 84, 86, 88, 140, 150, 152, 195; Cluster III (pink): residues 33, 35, 38, 41, 48, 52, 123, 125, 136, 138, 165, 171, 173, 199, 201. The hydrophilic conserved residues (dark blue): Asp 60, Asp 85, Asn 90, Gly 109, Lys 203. Conserved residues in the ligand-binding site (light blue): His 145, Tyr 192, or close by: Ala 87. Very few conserved residues are at the surface. Residues conserved between pentameric LGICs but not AChBP are indicated by yellow main chain (14, 19, 170 and the Cys-loop 124–135).

The high level of structural conservation, however, implies that the same topological regions form these interface contacts in all superfamily members (Fig. 5b). But in these interfaces different combinations of subunits will have different interactions, possibly creating variations in the precise allosteric contacts and movements in the various subclasses of these ion channels.

Activation mechanism

The location of the ligand-binding site is conserved among pentameric LGIC receptors¹ but the actual ligand-binding residues vary, creating specificity for different ligands. However, all ligand-gated ion channel LBDs have intrinsically the same function: the activation of a membrane pore. Indeed, a nicotinic receptor LBD is capable of activating a serotonin membrane channel in a chimaeric receptor⁴⁰. Thus, the essential activation mechanism is conserved across the superfamily. One option is that activation takes place by direct rotation of the protomer through a pivoting point at the ligand-binding site. But for any other mechanism, the location of the conserved residues implies that transmission of a signal from the binding pocket to the transmembrane domain takes place within the protomer and not through the interface region.

Within the protomer, several possible regions could transmit activation signals: changes in loop C, induced by ligand binding, could be transmitted directly into the transmembrane domain through the $\beta 9$ – $\beta 10$ β -hairpin; alternatively, large structural changes in the β -sheet regions could be transmitting the signal through the conserved Cys loop, acting directly on the membrane part of the pentameric LGICs. The movement observed at 9 Å for the *Torpedo* nAChR upon agonist binding¹⁹ fits well with the latter suggestion. We see a twisted β -sandwich, with two distinct hydrophobic cores, and it is possible that these cores move with respect to each other upon ligand binding. The effect of any of these activation mechanisms in the protomer will then be modulated by the varying subunit interfaces in the different subtypes of the receptor, allowing intricate specificity in neuronal signal transmission.

Conclusions

This crystal structure shows that molluscan AChBP is a homologue of the pentameric LGIC superfamily ligand-binding domains. It confirms the predicted Ig-topology, the location of the binding site at the subunit interface, the position of the MIR and the extensive data on the nicotinic ligand-binding residues.

With this structure, we present the first detailed three-dimensional information about the fold and the arrangement of the nicotinic ligand-binding site. It also clarifies the arrangement of subunits in nicotinic muscle receptors by showing the relative positioning of the principal and complementary part of the ligand-binding site.

The AChBP crystal structure provides an explanation for the role of the pentameric LGIC superfamily conserved residues. They stabilize the protomer structure by the formation of hydrophobic cores and packing of secondary structure elements. The crystal structure clarifies how the LGIC pentamers are built up, and how weakly the pentamer interfaces are conserved between LGICs, which has implications for the modes of action.

It is unclear whether AChBP performs the allosteric and desensitization movements that are important for pentameric LGIC function. AChBP is a soluble protein with a natural role in the modulation of synaptic transmission and lacks a transmembrane channel²⁷. Thus, ligand binding does not result in opening of a pore in AChBP, although some of the movements could be conserved in its evolution. Similarly, it is unclear whether non-competitive antagonist-binding sites are present in AChBP. Such questions can be addressed with chimaeric molecules and possibly by further crystallographic studies.

This structure will be highly relevant for the numerous drug-design studies that are targeting the pentameric LGIC superfamily.

The general structural knowledge on its folding will be applicable to the GABA, serotonin (5-HT₃) and glycine receptor fields. It will help us to understand their ligand-binding characteristics and hence could have an impact on the development of, for example, anti-emetics aimed at the 5-HT₃ receptor or the mood-defining drugs that target the GABA receptors. The availability of a three-dimensional description of the nicotinic ligand-binding site will be especially relevant for the design of new drugs against, for example, Alzheimer's disease, epilepsy and addiction to smoking. □

Methods

Protein purification

AChBP was cloned into expression vector pPIC9 (without residue Leu 1) and over-expressed in yeast, *Pichia pastoris* GS115, according to the Invitrogen manual. After 4 days of induction the medium was collected, concentrated and dialysed against standard buffer (20 mM Tris–HCl buffer (pH 8.0), 150 mM NaCl and 0.02% (w/v) NaN₃). The protein was purified by anion exchange (Poros50 HQ, MonoQ), and gel filtration (Superdex 200). It was dialysed against 50 mM HEPES buffer (pH 7.0) with 0.02% NaN₃ and concentrated to ~20 mg ml⁻¹. N-terminal sequencing revealed that part of the pPIC9-encoded signal sequence is retained, before residue 2 (sequence EAEAYVEF). The experimental relative molecular mass was 26,544 (MALDI), ~2K more than the calculated mass based on the sequence (24,649), owing to glycosylation at position Asn 66, as confirmed by deglycosylation experiments (data not shown).

Crystallization

We grew the crystals at room temperature using the hanging-drop vapour diffusion technique. All drops contained 2 µl of protein (10 mg ml⁻¹) and 2 µl of reservoir solution (9–11% (w/v) PEG 4000, 100 mM HEPES (pH 7.0), 50–200 mM CaCl₂ and 0.02% NaN₃). Depending on the batch of protein and the CaCl₂ concentration, we obtained three crystal forms. Orthorhombic and monoclinic crystals appeared under high CaCl₂ concentration and were frequently twinned. The orthorhombic crystals (P2₁2₁2₁) have the following cell constants: $a = 120.6$ Å, $b = 137.0$ Å, $c = 161.5$ Å, with two pentamer molecules per asymmetric unit (asu). The monoclinic crystals (P2₁) are very similar in morphology to the orthorhombic ones, but gave lower resolution data (~3.3 Å), with the following cell constants: $a = 121.1$ Å, $b = 162.1$ Å, $c = 139.4$ Å, $\beta = 90.13^\circ$, and four pentamers per asu. The tetragonal crystal form (P4₂2₁2) was obtained from a solution containing 11.5 (w/v) PEG 4000, 100 mM HEPES (pH 7.0), 150 mM CaCl₂ and 0.02% (w/v) NaN₃. They have the following cell constants: $a = b = 141.66$ Å, $c = 120.83$ Å and one pentamer per asu. For MAD experiments, orthorhombic and monoclinic crystals were soaked in mother liquor solution containing 5 mM and 10 mM trimethyl-lead acetate (MePb) respectively for 5 days. Before flash-cooling, all crystals were gradually equilibrated against mother liquor with 30% glycerol.

In all three space groups AChBP forms a decamer structure with 52 symmetry, where 2 pentamers have contact with each other through a calcium-binding site, at the 'top' of the $\alpha 1$ helix. This binding site (Asp 2 and Asp 5 from two protomers) is not conserved in the pentameric LGIC family. In the tetragonal space group the two-fold of the decamer coincides with a crystallographic two-fold. In solution, the AChBP protein behaves as a pentamer²⁷.

Structure determination and refinement

We collected native data (X11) and the Pb-1 data sets (BW7A) at the EMBL outstation at the DESY synchrotron in Hamburg and the Pb-2 data sets (BM14) at the ESRF, Grenoble (Table 1). Data were processed with DENZO/SCALEPACK⁴¹ (native) or MOSFLM⁴²/SCALA⁴³ software (Pb-1, Pb-2). The Pb sites, located at the interface of two pentamers, were found for both MAD sets by SOLVE⁴⁴ and heavy atom parameters were optimized with SHARP⁴⁵. NCS operators were found and refined with NCS6D and IMP⁴⁶. Multi-crystal averaging was executed with DM-Multi⁴⁷ using amplitudes from the monoclinic, orthorhombic and native (tetragonal) data sets, and experimental phases from the orthorhombic and monoclinic MAD experiments. The model was built in O⁴⁸ and refined with CNS⁴⁹, against the tetragonal 2.7 Å data. Refinement included partial fivefold NCS restraints, an overall anisotropic B factor and bulk solvent correction. The unusual vicinal disulphide bridge⁸, Cys 187–Cys 188, was visible in the electron density, but because of the limited resolution it was not analysed in detail. The final model contains 1,025 residues of AChBP pentamer, 5 HEPES molecules, 10 Ca²⁺ ions and 15 water molecules. The entire AChBP pentamer is well ordered, except for the N-terminus seven residues (part of the signal sequence) and the last five C-terminal residues. In addition, the HEPES, loop region 155–160 and the sugar residues attached to residue Asn 66 are not well resolved in the electron density. Root-mean-square deviations from ideal geometry for bond distances and angles are 0.01 Å and 1.6°, respectively.

Received 4 January; accepted 21 March 2001.

- Ortells, M. O. & Lunt, G. G. Evolutionary history of the ligand-gated ion-channel superfamily of receptors. *Trends Neurosci.* **18**, 121–127 (1995).
- Corring, P. J., Le Novère, N. & Changeux, J. P. Nicotinic receptors at the amino-acid level. *Annu. Rev. Pharmacol. Toxicol.* **40**, 431–458 (2000).
- Changeux, J. P. & Edelstein, S. J. Allosteric receptors after 30 years. *Neuron* **21**, 959–980 (1998).
- Arias, H. R. Localization of agonist and competitive antagonist binding sites on nicotinic acetylcholine receptors. *Neurochem. Int.* **36**, 595–645 (2000).

5. Paterson, D. & Nordberg, A. Neuronal nicotinic receptors in the human brain. *Prog. Neurobiol.* **61**, 75–111 (2000).
6. Galzi, J. L. *et al.* Identification of a novel amino acid alpha-tyrosine 93 within the cholinergic ligands-binding sites of the acetylcholine receptor by photoaffinity labeling. Additional evidence for a three-loop model of the cholinergic ligands-binding sites. *J. Biol. Chem.* **265**, 10430–10437 (1990).
7. Dennis, M. *et al.* Amino acids of the *Torpedo marmorata* acetylcholine receptor alpha subunit labeled by a photoaffinity ligand for the acetylcholine binding site. *Biochemistry* **27**, 2346–2357 (1988).
8. Kao, P. N. & Karlin, A. Acetylcholine receptor binding site contains a disulfide cross-link between adjacent half-cystinyl residues. *J. Biol. Chem.* **261**, 8085–8088 (1986).
9. Middleton, R. E. & Cohen, J. B. Mapping of the acetylcholine binding site of the nicotinic acetylcholine receptor: [3H]nicotine as an agonist photoaffinity label. *Biochemistry* **30**, 6987–6997 (1991).
10. Fu, D. X. & Sine, S. M. Competitive antagonists bridge the alpha-gamma subunit interface of the acetylcholine receptor through quaternary ammonium-aromatic interactions. *J. Biol. Chem.* **269**, 26152–26157 (1994).
11. O'Leary, M. E., Filatov, G. N. & White, M. M. Characterization of *d*-tubocurarine binding site of *Torpedo* acetylcholine receptor. *Am. J. Physiol.* **266**, C648–653 (1994).
12. Corringer, P. J. *et al.* Identification of a new component of the agonist binding site of the nicotinic alpha 7 homo-oligomeric receptor. *J. Biol. Chem.* **270**, 11749–11752 (1995).
13. Sine, S. M., Kreienkamp, H. J., Bren, N., Maeda, R. & Taylor, P. Molecular dissection of subunit interfaces in the acetylcholine receptor: identification of determinants of alpha-conotoxin M1 selectivity. *Neuron* **15**, 205–211 (1995).
14. Czajkowski, C., Kaufmann, C. & Karlin, A. Negatively charged amino-acid residues in the nicotinic receptor delta subunit that contribute to the binding of acetylcholine. *Proc. Natl Acad. Sci. USA* **90**, 6285–6289 (1993).
15. Martin, M., Czajkowski, C. & Karlin, A. The contributions of aspartyl residues in the acetylcholine receptor gamma and delta subunits to the binding of agonists and competitive antagonists. *J. Biol. Chem.* **271**, 13497–13503 (1996).
16. Prince, R. J. & Sine, S. M. Molecular dissection of subunit interfaces in the acetylcholine receptor. Identification of residues that determine agonist selectivity. *J. Biol. Chem.* **271**, 25770–25777 (1996).
17. Dougherty, D. A. Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science* **271**, 163–168 (1996).
18. Unwin, N. Nicotinic acetylcholine receptor at 9 Å resolution. *J. Mol. Biol.* **229**, 1101–1124 (1993).
19. Unwin, N. Acetylcholine receptor channel imaged in the open state. *Nature* **373**, 37–43 (1995).
20. Beroukhi, R. & Unwin, N. Three-dimensional location of the main immunogenic region of the acetylcholine receptor. *Neuron* **15**, 323–331 (1995).
21. Miyazawa, A., Fujiyoshi, Y., Stowell, M. & Unwin, N. Nicotinic acetylcholine receptor at 4.6 Å resolution: transverse tunnels in the channel wall. *J. Mol. Biol.* **288**, 765–786 (1999).
22. West, A. P. Jr, Bjorkman, P. J., Dougherty, D. A. & Lester, H. A. Expression and circular dichroism studies of the extracellular domain of the alpha subunit of the nicotinic acetylcholine receptor. *J. Biol. Chem.* **272**, 25468–25473 (1997).
23. Schratzenholz, A. *et al.* Expression and renaturation of the N-terminal extracellular domain of *Torpedo* nicotinic acetylcholine receptor alpha-subunit. *J. Biol. Chem.* **273**, 32393–32399 (1998).
24. Alexeev, T. *et al.* Physicochemical and immunological studies of the N-terminal domain of the *Torpedo* acetylcholine receptor alpha-subunit expressed in *Escherichia coli*. *Eur. J. Biochem.* **259**, 310–319 (1999).
25. Wells, G. B., Anand, R., Wang, F. & Lindstrom, J. Water-soluble nicotinic acetylcholine receptor formed by alpha7 subunit extracellular domains. *J. Biol. Chem.* **273**, 964–973 (1998).
26. Tierney, M. L. & Unwin, N. Electron microscopic evidence for the assembly of soluble pentameric extracellular domains of the nicotinic acetylcholine receptor. *J. Mol. Biol.* **303**, 185–196 (2000).
27. Smit, A. B. *et al.* A glia-derived acetylcholine-binding protein that modulates synaptic transmission. *Nature* **411**, 261–268 (2001).
28. Bork, P., Holm, L. & Sander, C. The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.* **242**, 309–320 (1994).
29. Le Novère, N., Corringer, P. J. & Changeux, J. P. Improved secondary structure predictions for a nicotinic receptor subunit: incorporation of solvent accessibility and experimental data into a two-dimensional representation. *Biophys. J.* **76**, 2329–2345 (1999).
30. Holm, L. & Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **25**, 231–234 (1997).
31. Tzartos, S. J. *et al.* The main immunogenic region (MIR) of the nicotinic acetylcholine receptor and the anti-MIR antibodies. *Mol. Neurobiol.* **5**, 1–29 (1991).
32. Fernando Valenzuela, C., Weign, P., Yguerabide, J. & Johnson, D. A. Transverse distance between the membrane and the agonist binding sites on the *Torpedo* acetylcholine receptor: a fluorescence study. *Biophys. J.* **66**, 674–682 (1994).
33. Galzi, J. L., Bertrand, S., Corringer, P. J., Changeux, J. P. & Bertrand, D. Identification of calcium binding sites that regulate potentiation of a neuronal nicotinic acetylcholine receptor. *EMBO J.* **15**, 5824–5832 (1996).
34. Machold, J., Weise, C., Utkin, Y., Tsetlin, V. & Hucho, F. The handedness of the subunit arrangement of the nicotinic acetylcholine receptor from *Torpedo californica*. *Eur. J. Biochem.* **234**, 427–430 (1995).
35. Zhong, W. *et al.* From ab initio quantum mechanics to molecular neurobiology: a cation-pi binding site in the nicotinic receptor. *Proc. Natl Acad. Sci. USA* **95**, 12088–12093 (1998).
36. Harel, M. *et al.* Quaternary ligand binding to aromatic residues in the active-site gorge of acetylcholinesterase. *Proc. Natl Acad. Sci. USA* **90**, 9031–9035 (1993).
37. Hemmingsen, J. M., Gernert, K. M., Richardson, J. S. & Richardson, D. C. The tyrosine corner: a feature of most Greek key beta-barrel proteins. *Protein Sci.* **3**, 1927–1937 (1994).
38. Mishina, M. *et al.* Location of functional regions of acetylcholine receptor α -subunit by site-directed mutagenesis. *Nature* **313**, 364–369 (1985).
39. Green, W. N. & Wanamaker, C. P. Formation of the nicotinic acetylcholine receptor binding sites. *J. Neurosci.* **18**, 5555–5564 (1998).
40. Eisele, J. L. *et al.* Chimaeric nicotinic-serotonergic receptor combines distinct ligand binding and channel specificities. *Nature* **366**, 479–483 (1993).
41. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
42. Leslie, A. G. Integration of macromolecular diffraction data. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 1696–1702 (1999).
43. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
44. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 849–861 (1999).
45. La Fortelle, E. de & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement in the MIR and MAD methods. *Methods Enzymol.* **276**, 472–494 (1997).
46. Kleywegt, G. J. & Jones, T. A. in *From First Map to Final Model* (eds Bailey, S., Hubbard, R. & Waller, D.) 59–66 (SERC Daresbury Laboratory, Warrington, 1994).
47. Cowtan, K. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* **31**, 34–38 (1994).
48. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
49. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
50. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882 (1997).

Acknowledgements

We thank R. van der Schors for mass spectrometry; R. van Elk for HEPES binding assays; A. Perrakis and U. Gohlke for helpful suggestions; M. Lamers for help with figure preparation; and beam line scientists at ESRF and EMBL outstations Hamburg and Grenoble, in particular W. Rypniewski and G. Leonard for assistance during data collection. Nederlandse Organisatie voor Wetenschappelijk Onderzoek-Chemische Wetenschappen and Stichting voor de Technische Wetenschappen are acknowledged for financial support.

Correspondence should be addressed to T.K.S. (e-mail: sixma@nki.nl). The structure has been deposited in the PDB with ID number 119B.

Spherical episodic ejection of material from a young star

J. M. Torrelles*, N. A. Patel†, J. F. Gómez‡, P. T. P. Ho†, L. F. Rodríguez§, G. Anglada||, G. Garay¶, L. Greenhill†, S. Curiel☆ & J. Cantó☆

* Institut d'Estudis Espacials de Catalunya (IEEC/CSIC) and Instituto de Ciencias del Espacio (CSIC), Gran Capità 2, 08034 Barcelona, Spain

† Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

‡ Laboratorio de Astrofísica Espacial y Física Fundamental (INTA), Apdo. Correos 50727, 28080 Madrid, Spain

§ Instituto de Astronomía (UNAM), Apdo. Postal 72-3, 58089 Morelia, México

|| Instituto de Astrofísica de Andalucía (CSIC), Ap. Correos 3004, 18080 Granada, Spain

¶ Departamento de Astronomía, Universidad de Chile, Casilla 36-D, Santiago, Chile

☆ Instituto de Astronomía (UNAM), Apdo. Postal 70-264, DF 04510, México

The exact processes by which interstellar matter condenses to form young stars are of great interest, in part because they bear on the formation of planets like our own from the material that fails to become part of the star. Theoretical models suggest that ejection of gas during early phases of stellar evolution is a key mechanism for removing excess angular momentum, thereby allowing material to drift inwards towards the star through an accretion disk^{1,2}. Such ejections also limit the mass that can be accumulated by the stellar core^{1,2}. To date, these ejections have been observed to be bipolar and highly collimated, in agreement with theory. Here we report observations at very high angular resolution of the proper motions of an arc of water-vapour masers near a very young, massive star in Cepheus. We find that the arc of masers can be fitted to a circle with an accuracy of one part in a thousand, and that the structure is expanding. Only a sphere will always produce a circle in projection, so our observations strongly suggest that the perfectly spherical ejection of material from this star took place about 33 years earlier. The spherical symmetry of the ejecta and its episodic nature are very surprising in the light of present theories.

Here we make use of the nature of the 1.35-cm rotational transition of H₂O as a maser (microwave amplification by stimulated emission of radiation), with brightness temperatures exceeding 10¹⁰ K³. The high brightness and the compact nature of masers have proven to be extremely useful in astrophysical very-long-baseline interferometry (VLBI) studies⁴. The very high angular resolution of VLBI allows proper motion studies of masers on the timescales of weeks. This has led to important discoveries, including a new way to measure celestial distances and extremely tight constraints on the mass of supermassive black holes in extragalactic nuclei^{5,6}.

The star-forming region known as Cepheus A is the second source in the sky ever noted to exhibit the phenomenon of bipolar molecular outflow^{7,8}. The driver of the outflow was soon identified as a radio continuum source associated with a high-mass star deeply embedded in high-density molecular gas, which makes it invisible at optical and even at infrared wavelengths. Upon closer study it was revealed to be a biconical radio jet, although its expected associated circumstellar disk was not seen with thermal molecular lines, probably because of their expected faintness^{9–13}. However, fortunately, strong H₂O masers are present in this region. Through an experiment on the very large array (VLA) we showed that there is a distribution of masers localized spatially in the form of a flattened disk, centred on the jet, and perpendicular to it¹¹. Such a configuration for the disk and the jet is as predicted by theory, because it is the accretion process within the disk which drives the outflow process in the standard model¹.

Using the very long baseline array (VLBA) of the National Radio Astronomy Observatory, we obtained data during three epochs in 1996 in order to study the proper motions of the masers. With an improvement in angular resolution by a factor of 180, we find that most of the unresolved maser features seen on the VLA are now resolved into linear or arclike microstructures (0.4–70 AU).

The detection of the proper motions of these masers depends on a number of considerations: (1) Very accurate relative positions of individual maser features must be achieved, as a proper motion of 10 km s⁻¹ at a distance of 725 pc yields an angular displacement of only 3 milliarcseconds (mas) in a year. This is accomplished by phase referencing to a single isolated maser feature. The relative positional accuracy is then a very small fraction of the synthesized beam size, on the order of 50 microarcseconds. (2) The maps at each epoch must be registered relative to each other. Because any individual maser spot can have its own proper motion, referencing to a single maser spot can result in arbitrary offsets in proper motions between epochs. We avoid this problem by using the average position of a number of strong maser spots (spread over a region of about 5 arcsec size) to define a stationary reference position. Although this implicitly assumes that the individual proper motions of the individual maser spots average to zero, we find that this technique in fact works well. In our field of view, we determined proper motions which point in different directions, and which are furthermore systematically perpendicular to the curvatures of the maser structures. This would not be possible if there were a significant component of motion which has been introduced by the use of a nonstationary reference position. (3) Whereas individual parts of a microstructure of masers appear and disappear, the arclike structures can be recognized easily from epoch to epoch. This allowed us to visualize and measure their proper motions as an entire structure.

Here we concentrate on only one particular and most unusual maser feature. In Fig. 1A we show the arcuate structure of masers (hereafter called R5) found south of Cepheus A HW2, at a relatively large distance (500 AU) from the circumstellar disk associated with this young stellar object (YSO). The detected structure, at 100 mas in length, has an angular size similar to that of the VLA beam of the previous observations¹¹, and is large enough to show up as extended emission in the corresponding VLA maps (Fig. 1B). In contrast, the R5 structure is 200 times the synthesized angular resolution achieved with the VLBA. Such 'large' coherent structures, embedded within a large (5 arcsec) field of view, would have been difficult to map in the early days of VLBI because of the large field of view which must be synthesized, and this has only recently become possible because of the rapid improvements in computing power.

Figure 1C shows that the arc of masers is fitted extremely well by a circle of radius 62 AU. The deviation from the circle is at the level of 0.1% for all three epochs. The proper motions of individual sections of the arc are consistent with uniform expansion perpendicular to the arc at the rate of 9 km s⁻¹ with a dynamical timescale of 33 yr (see Fig. 1A). The high degree of symmetry exhibited both by the spatial structure and by the proper motions suggests that we must be seeing the limb-brightened parts of a spherical structure. In particular, the high degree of conformity to a circle makes it unlikely that other structures such as rings or disks or any other geometry might work, because they must lie exactly in the plane of the sky. That the masers are all limb-brightened structures is also consistent with the fact that their line-of-sight local standard of rest (LSR) velocities are nearly the same, meaning that their full three-dimensional proper motions are in fact perpendicular to the line of sight. The lack of masers off the arc suggests that the structure must be an extremely thin shell. That the masers are all located within one quadrant of a circle may suggest that there is another important process that can modify this highly symmetrical outflow. This could be local inhomogeneity or the winds of another source nearby, such as Cepheus A HW2.

We note that this arc structure consists of very small individual

pieces (Fig. 1A), each roughly linear in shape (length sizes of 0.4–1 AU) but being unresolved in the perpendicular direction, and each with a coherent velocity structure which can be described as a single spectral line or feature. The maser emissions in the five subregions indicated in Fig. 1A (labelled a–e) have almost the same radial velocities during the three epochs, with $V_{\text{LSR}} = -8.7 \text{ km s}^{-1}$ (a, c, e), $V_{\text{LSR}} = -9.6 \text{ km s}^{-1}$ (b) and $V_{\text{LSR}} = -8.5 \text{ km s}^{-1}$ (d), (systemic velocity $= -11.7 \text{ km s}^{-1}$)¹¹ and line widths $\leq 0.7 \text{ km s}^{-1}$. The flat-

tened structures and their motions perpendicular to the arc strongly suggest formation and excitation by a shock. The size scale of the individual pieces may be a measure of the thickness of the shell itself. The broken appearance of the masers along the arc is probably due to the nonthermal nature of the masering process as well as instabilities within the thin shell. These data can be seen to support the expectations from theoretical models of shock excitation of water masers^{14,15}.

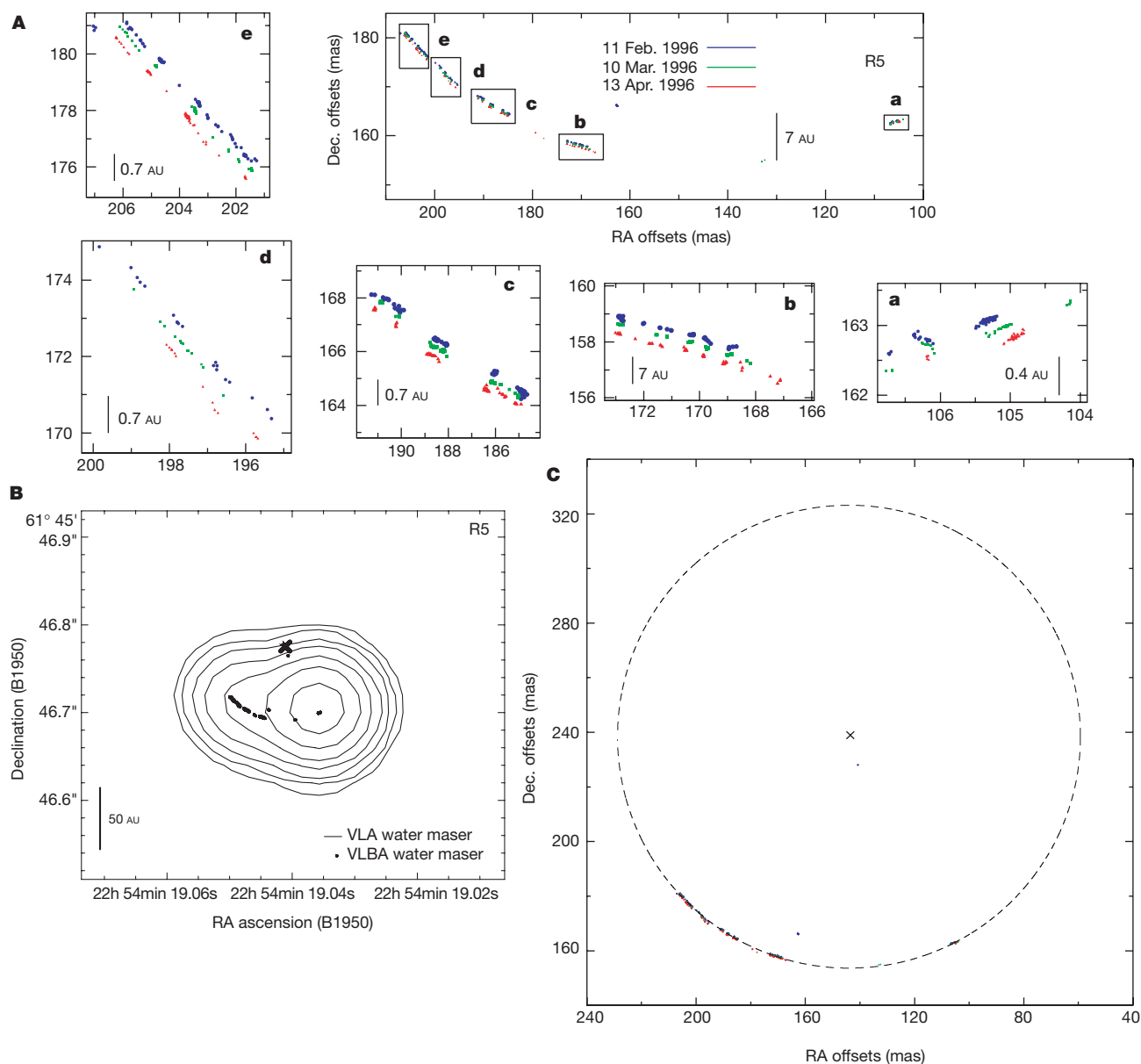


Figure 1 Water maser emission near the young star Cepheus A HW2. The H_2O maser emission ($6_{16}-5_{23}$ line at 22.235 GHz) was observed with the VLBA at an angular resolution of 0.5 mas and a velocity resolution of 0.21 km s^{-1} . We simultaneously mapped for each epoch around 60 fields of 512×512 pixels $\times$ 245 velocity channels with cell size 0.06 mas. Most of the masers detected with the VLA seven months earlier were detected with the VLBA. We focus on one of the maser features (R5) detected 0.7 arcsec (500 AU) south of Cepheus A HW2, which has been resolved into an arc of 100 mas (72 AU at the distance of 725 pc) in length. **A**, Water maser positions measured with the VLBA. Circles (blue), squares (green), and triangles (red) correspond to the masers detected on 11 February, 10 March and 13 April 1996, respectively. Close-ups of five subregions (labelled a–e) are also shown. Offset positions are relative to a reference position located at $\alpha(\text{J2000}) = 22^{\text{h}} 56^{\text{m}} 17.967^{\text{s}}$, $\delta(\text{J2000}) = 62^{\circ} 01' 48.75''$. **B**, Contour map of

the water maser emission seen with the VLA on 5 July 1995 towards the region R5 at $V_{\text{LSR}} = -8.9 \text{ km s}^{-1}$. Contour levels are 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, $0.64 \times 603 \text{ Jy beam}^{-1}$ (VLA beam size is 80 mas)¹¹. Dots indicate the positions of the masers constituting the arc microstructure R5 as measured with the VLBA. The cross marks the centre of the circle which is fitted to the arc of masers in the next panel. **C**, The arc structure is fitted by least squares to a circle (including the data of all three epochs). The cross indicates the position of the centre of the circle. Each epoch can be fitted to its own circle to a precision of 0.1%. The radius of the fitted circles is increasing with time, consistent with spherical expansion. There are two maser spots, detected only during the first epoch, which do not fall on the circle. They have radial velocities which disagree with the radial velocity on the circle, suggesting that these are spots which may be lying on the front surface of the spherical shell.

The coherent velocity structure within the arc is seen in Fig. 2, which shows a position–radial velocity (LSR) map of the maser spots in subregion (a) for all three epochs. We note in Fig. 2 that for individual features, for example at distances of 0.3–1.8 mas and $V_{\text{LSR}} = -8.5$ to -9.5 km s^{-1} , we can see a clear velocity trend within our synthesized angular resolution of 0.5 mas. This velocity trend persists in all three epochs. Such velocity gradients are probably an indication of a shape or local density gradient so that the shock surface is not completely flat. A curvature in the shock surface, as in a small bow shock structure, results in a spread in projected radial velocity. That the very large velocity wings visible in certain cases seem to be well localized in space, in contrast to the observable velocity gradients in the core of the line, may be a manifestation of the maser amplification process, which favours the line of sight with just the right amplification conditions, and the fact that the shell is very thin. In addition, all of these linear microstructures (a–e) are tangential to the arc curvature, whereas the whole curved structure shows a spatial displacement predominantly perpendicular to the surface of the arc structure (see Fig. 1A).

The observed proper motions and morphology of the arc structure as a whole suggest that it represents a spherical bubble, probably driven by a YSO located at the centre of the circle which fits the arc microstructure. To date, no known continuum source at centimetre wavelengths has been reported at this position¹². However, since our recognition of this circle of masers, we have examined archival data from the VLA that indicate that a very faint source at 3.6 cm at the level of 0.2 mJy (epoch 1991) may have been present. Further observations with an angular resolution better than tenths of an arcsecond will be important for separating any underlying source from other continuum sources in the region. Submillimetre observations of the dust may also be important if this new centre of star formation activity is deeply embedded. We speculate that the underlying star is probably not a late-type star which typically

has a much more irregular shape when observed in water maser emission¹⁶.

The extremely uniform shape and thinness of the R5 arc suggest that the observed material must derive from a single ejection event of short duration. For the same reasons, it is also likely that the masering material is not from local material which is being swept but rather is direct ejecta from the young star itself. Otherwise, local interactions in the sweeping process would not result in such a coherent structure at the 0.1% level. Nevertheless, there are local structural perturbations and large velocity dispersions in the arc structure. If these local differences are present in the initial ejection, again, we would expect that the overall morphology would not be coherent by the time the structure has reached its present size. We speculate that a possible scenario is an initially smoothly expanding shell which is now overtaken by a second episode of ejection. The strongest support for this conjecture may be the very large spread in velocity seen in the line wings in some of the spots which remain localized within the arc structure.

The detection of maser structures organized into arcs and other shapes, both on large (0.1 arcsec) and small (1 mas) scales, have also been reported previously^{17–19}. Some of the more recent studies such as the remarkable microstructures in S106IR²⁰ (bow-shock) and IRAS 21391+5802²¹ (loop) also support shock excitation. In particular, the S106IR bow-shock appears to be created by the impact of a very collimated jet with ambient matter. However, all of these maser structures are typically very irregular in shape and kinematics. The uniqueness of our results in R5 lies in the extreme circular symmetry and apparent smoothness of the structure and the ability to trace a very large fraction of the circle. These results are perhaps the best evidence yet for masers tracing a shock front. In addition, the remarkable linear ‘building blocks’ forming the R5 arc structure suggest that although the structure of the shock front depends on the local conditions, the shock front itself is continuous and organized over a very large solid angle as subtended from the exciting source. Symmetrical structures can be distorted easily by any number of means including ambient density structures or magnetic fields, but an inherently asymmetrical structure cannot easily be made to appear symmetrical. These results suggest that the ejection process in young stellar systems, at least in this case, is highly spherically symmetric, short-lived, and episodic in the beginning. □

Received 29 December 2000; accepted 21 March 2001.

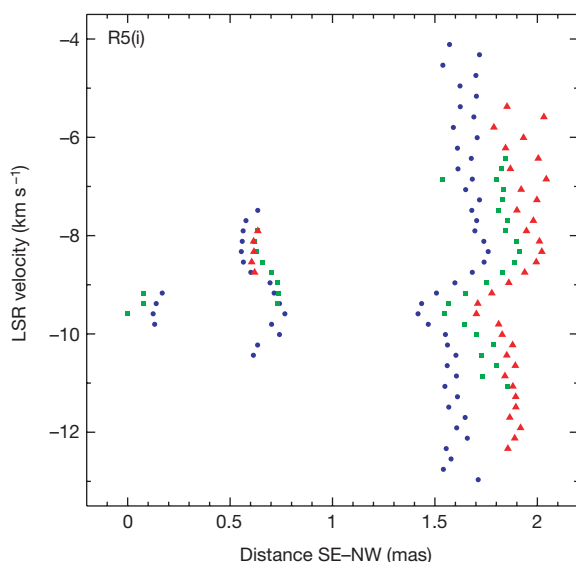


Figure 2 Tracking the velocity structure of the maser spots over two months. This position–velocity (LSR) map of the maser spots in subregion (a) of the arc structure (Fig. 1A) for the three epochs (circles, squares and triangles respectively), has been obtained along an axis with PA = -68° , centred on the easternmost maser spot. Velocity structures can be seen to persist in time and are stable in nature. The plot shows that individual maser spots as well as the arc structures can be easily traced from epoch to epoch. The evolution of these velocity structures as well as the circular morphology of the shell of masers will be very interesting problems to study in the future. Given the short dynamical age of 33 yr, we expect that significant changes can be seen in future proper-motion VLBI studies.

- Shu, F. H., Adams, F. C. & Lizano, S. Star formation in molecular clouds—Observations and theory. *Annu. Rev. Astron. Astrophys.* **25**, 23–81 (1987).
- Lizano, S. & Torrelles, J. M. (eds) Circumstellar disks, outflows, and star formation. *Rev. Mex. Astron. Astrofis. Ser. Conf.* **1**, (1995).
- Cheung, A. C., Rank, D. M., Townes, C. H., Thornton, D. D. & Welch, W. J. Detection of water in interstellar regions by its microwave radiation. *Nature* **221**, 626–628 (1969).
- Reid, M. J. & Moran, J. M. Masers. *Annu. Rev. Astron. Astrophys.* **19**, 231–276 (1981).
- Genzel, R., Reid, M. J., Moran, J. M. & Downes, D. Proper motions and distances of H₂O maser sources. I—The outflow in Orion-KL. *Astrophys. J.* **244**, 884–902 (1981).
- Miyoshi, M. *et al.* Evidence for a black hole from high-rotation velocities in a sub-parsec region of NGC4258. *Nature* **373**, 127–129 (1995).
- Rodríguez, L. E., Ho, P. T. P. & Moran, J. M. Anisotropic mass outflow in Cepheus A. *Astrophys. J.* **240**, L149–L152 (1980).
- Ho, P. T. P., Moran, J. M. & Rodríguez, L. F. Mass outflow in star formation regions—Cepheus A. *Astrophys. J.* **262**, 619–635 (1982).
- Torrelles, J. M., Rodríguez, L. F., Cantó, J. & Ho, P. T. P. A circumstellar molecular gas structure associated with the massive young star Cepheus A-HW 2. *Astrophys. J.* **404**, L75–L78 (1993).
- Rodríguez, L. F. *et al.* Cepheus A HW2: A powerful thermal radio jet. *Astrophys. J.* **430**, L65–L68 (1994).
- Torrelles, J. M. *et al.* The thermal radio jet of Cepheus A HW2 and the water maser distribution at 0.08″ scale (60 AU). *Astrophys. J.* **457**, L107–L111 (1996).
- Torrelles, J. M. *et al.* Systems with H₂O maser and 1.3 centimeter continuum emission in Cepheus A. *Astrophys. J.* **509**, 262–269 (1998).
- Hughes, V. A. & Wouterloot, J. G. A. The star-forming region in Cepheus A. *Astrophys. J.* **276**, 204–210 (1984).
- Elitzur, M., Hollenbach, D. J., McKee, C. F. Planar H₂O masers in star-forming regions. *Astrophys. J.* **394**, 221–227 (1992).
- Kaufmann, M. J. & Neufeld, D. A. Water maser emission from magnetohydrodynamic shock waves. *Astrophys. J.* **456**, 250–263 (1996).
- Marvel, K. B. The circumstellar environment of evolved stars as revealed by studies of circumstellar water masers. *Publ. Astron. Soc. Pacif.* **109**, 1286–1287 (1997).

17. Gwinn, C. R. Hypersonic acceleration and turbulence of H₂O masers in W49N. *Astrophys. J.* **429**, 241–252 (1994).
18. Leppanen, K., Liljestrom, T. & Diamond, P. Submillisecond linear polarization observations of water masers in W51M. *Astrophys. J.* **507**, 909–918 (1998).
19. Claussen, M. J., Marvel, K. B., Wootten, A. & Wilking, B. A. Distribution and motion of the water masers near IRAS 05413-0104. *Astrophys. J.* **507**, L79–L82 (1998).
20. Furuya, R. S. *et al.* A microjet: a protostar's cry at birth. *Astrophys. J.* **542**, L135–L138 (2000).
21. Patel, N. A. *et al.* Proper motion of water masers associated with IRAS 21391+5802: Bipolar outflow and an AU-scale dusty circumstellar shell. *Astrophys. J.* **538**, 268–274 (2000).

Acknowledgements

NRAO is operated by Associated Universities, Inc., under cooperative agreement with the National Science Foundation. This work was partially supported by MCYT and Programa de Cooperación con Iberoamérica (Spain), and DGAPA and CONACyT (México).

Correspondence and requests for materials should be addressed to P.T.P.H. (e-mail: ho@cfa.harvard.edu).

Observation of the ideal Josephson effect in superfluid ⁴He

Kalyani Sukhatme*, Yury Mukharsky*†, Talso Chui* & David Pearson*

* Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA

† CEA-DRECAM, Service de Physique de l'État Condensé, Centre d'Études de Saclay, 91191 Gif-sur-Yvette Cedex, France

Superfluids and superconductors are the only states of condensed matter that can be described by a single wavefunction, with a coherent quantum phase Φ . The mass flow in a superfluid can be described by classical hydrodynamics for small flow velocity, but above a critical velocity, quantized vortices are created and the classical picture breaks down. This can be observed for a superfluid flowing through a microscopic aperture when the mass flow is measured as a function of the phase difference across the aperture; the curve resembles a hysteretic sawtooth where each jump corresponds to the creation of a vortex^{1–3}. When the aperture is made small enough, the system can enter the so-called 'ideal' Josephson regime^{1,4}, where the superfluid mass flow becomes a continuous function of the phase difference. This regime has been detected^{1,5,6} in superfluid ³He, but was hitherto believed to be unobservable, owing to fluctuations⁷, in ⁴He. Here we report the observation of the ideal Josephson effect in ⁴He. We study the flow of ⁴He through an array of micro-apertures and observe a transition to the ideal Josephson regime as the temperature is increased towards the superfluid transition temperature, T_λ .

In 1962, Josephson² predicted that if two superconductors are separated by a thin insulator, then the superposition of the wavefunctions of the bulk superconductors causes the current through the junction to depend sinusoidally on the difference in phase⁸, $I = I_c \sin(\phi)$. This equation is known as the d.c. Josephson relation. Here, ϕ is the phase difference ($\Phi_1 - \Phi_2$) between the two bulk systems. Similar current–phase relationships, $I(\phi)$, are observed in other superconductor/superfluid systems with a separating constriction/insulator (these systems are generally known as 'weak links'). Such systems exhibit a variety of $I(\phi)$ relationships of which non-hysteretic ones are referred to as demonstrating the ideal Josephson effect⁴. In order to observe the ideal Josephson effect, it is necessary for the dimensions of the weak link to be comparable to the minimal length over which the wavefunction can change (coherence length, ξ). The velocity of superflow, v_s , is proportional to the phase difference across the link. This phase difference can be changed by application of a pressure differential,

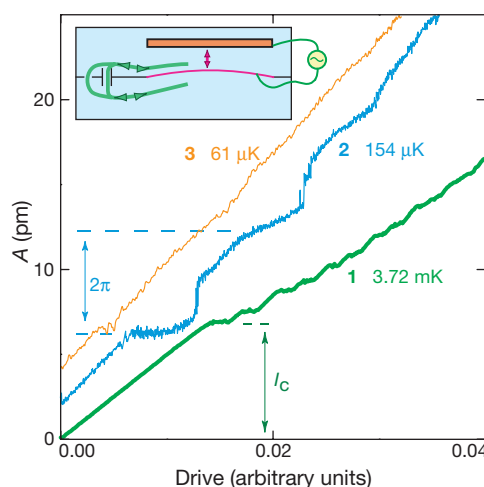


Figure 1 Staircase patterns (amplitude versus drive). $T_\lambda - T$ is 3.72 mK (**1**); 154 μ K (**2**); 61 μ K (**3**). The data for curve **1** has been divided by 20 in x and by 10 in y . The plots are shifted 2 pm for clarity. The height of the first step corresponds to the critical current I_c , as marked. Each subsequent step corresponds to an additional phase difference of 2π , as marked. The insert shows the schematics of the cell: it consists of a two-chamber resonator connected by a flow path. One of the walls between the two chambers is a flexible diaphragm made of thin metallized kapton. The diaphragm and the adjacent wall of the volume form a capacitor, which is used to apply an a.c. pressure drive to the resonator. The drive, in concert with the pressure difference exerted across the flow path by the diaphragm, generates a helium flow. Diaphragm motion is measured with a SQUID-based position sensor with a resolution of about 2×10^{-13} m Hz^{-1/2}. The calibration of this position sensor is difficult and thus the magnitudes of displacements or currents cited here should be taken as estimates only. The temperature of the cell is controlled using high-resolution thermometers¹⁵ with nano-Kelvin temperature stability.

ΔP , according to the a.c. Josephson–Anderson relation³ $\dot{\phi} = -(m_4/\hbar)(\Delta P/\rho)$ (ρ is the helium density, m_4 is the mass of a Helium atom). As ϕ increases, so does the velocity, until its critical value is reached. When the link dimension is much larger than the coherence length, the critical velocity corresponds to a phase difference greater than 2π . In this regime, the mass current, I , is also proportional to the phase difference, because the current density is given by $\rho_s v_s$, where ρ_s is the density of the superfluid fraction. Thus, the current–phase relationship is linear up to a critical velocity at which one or more 'phase slip' events⁵ occur, resulting in a decrease in the phase difference by an integral multiple of 2π . Each 'phase slip' reduces the current by a quantized amount and thus dissipates the kinetic energy of the flow. The system is said to be in the phase-slip regime. As the coherence length is increased (by approaching the superfluid transition, for example) relative to the link size, the critical velocity decreases and corresponds to a smaller phase difference. However, 2π phase slips cannot happen at a phase lower than π , because this would increase the energy. Instead, it becomes energetically favourable to reduce the amplitude of the wavefunction (proportional to ρ_s) within the link⁴ as the phase difference approaches π . The $I(\phi)$ becomes non-hysteretic (continuous). Here, the system is in the Josephson regime. By combining the d.c. and a.c. Josephson relations, one obtains $I = I_c \sin(\int - (m_4/\hbar)(\Delta P/\rho) dt)$. Thus, it is clear that both positive and negative values of the current can result from the application of a positive pressure head. We call this feature (which is very characteristic of the Josephson regime) the 'negative inertia' of the liquid. Another sign of entering the Josephson regime is the disappearance of the dissipation associated with the phase slips. Our experiments provide the first direct evidence of these two facts in ⁴He.

Electron-beam lithography techniques permit us to make structures with dimensions of the order of 0.1 μ m. This is comparable to

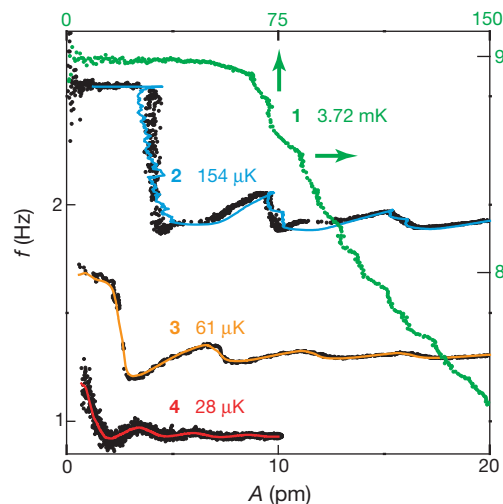


Figure 2 Frequency versus amplitude of oscillations. The curves are recorded by inserting the resonator into a phase-lock loop, which maintained a constant phase angle (90 degrees in most cases) between the applied drive and the resonator's response as the drive level was varied. In this scheme, the variations of the resonant frequency are tracked directly by the feedback loop. $T_\lambda - T$ is 3.72 mK (1); 154 μ K (2); 61 μ K (3); 28 μ K (4). The red and the orange lines are fits to the data (using eight terms in the Fourier expansion) as described in the text. At temperatures below about 100 μ K below transition, the fits become unsatisfactory (probably because in this temperature range a hysteretic form of $I(\phi)$, which cannot be approximated by a Fourier expansion, is expected). We assumed an $I(\phi)$ as shown in Fig. 3 (curve 2) and numerically simulated the response of our resonator. The result is shown in blue (curve 2). The magnitude of the fluctuations of the critical velocity has to be assumed to be zero to make the fit satisfactory. Curve 1 has been plotted, using scale on right-top axes. To enhance its features, curve 4 has been stretched vertically by a factor of 2 around its high-amplitude value.

the coherence length in ^3He and conventional superconductors, but is much larger than that of ^4He ($\xi_0 \approx 10^{-10}$ m) at low temperatures. As the temperature approaches T_λ , the coherence length diverges as $\xi = \xi_0(1 - T/T_\lambda)^{-\nu}$, with $\nu = 0.672$. Thus the temperature must approach T_λ to within about 100 μ K to achieve a coherence length of the order of 0.1 μ m. By varying the temperature, it is possible to go continuously from the phase-slip regime to the Josephson regime. However, in ^4He , fluctuations of the order parameter in a volume with dimensions of the order of ξ are predicted to be comparable to the order parameter itself⁷. This has led to conjecture that the ideal Josephson effect in ^4He would be smeared out and could not be observed. But in spite of these concerns, our data show that the Josephson signature is indeed seen, and the fluctuations do not smear it out.

Our experiments were performed in a Helmholtz resonator-like cell similar to that described in ref. 1. The schematic of the cell is shown in the inset to Fig. 1, along with a description. The Helmholtz resonator is driven to generate a pressure difference and a mass flow across the flow path. The pressure difference changes the quantum phase ϕ across the flow path according to the a.c. Josephson relation. The mass flow is calculated from the rate of change of the diaphragm position. Measurements were made in the temperature range: 80 mK $> (T_\lambda - T) > 20$ μ K.

The weak link consists of a rectangular array of 24 slit-apertures of dimensions 3 μ m by 0.17 μ m each, separated by about 10 μ m and made in a 0.15- μ m-thick membrane. An array of slit-apertures was chosen, in order to increase the total mass current through the weak link, while keeping the characteristic size of the link small. The array, together with a parallel macroscopic flow path, forms the inertial element of the resonator whose elastic element is the flexible diaphragm. Far from T_λ , the current through the flow path is

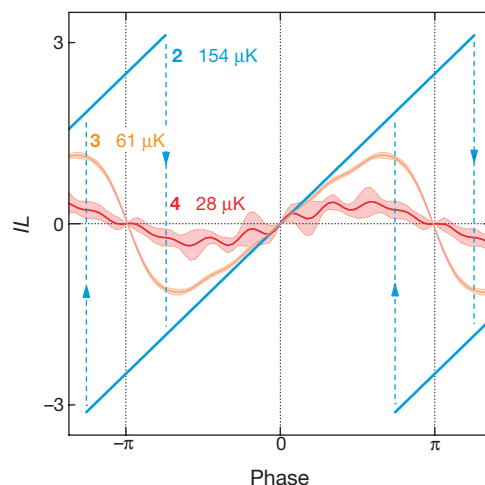


Figure 3 The current-phase characteristics. Curves 3 and 4 have been obtained from the fits shown in Fig. 2 (curves 3 and 4, respectively). 68% confidence intervals are shown. The hysteretic $I(\phi)$, curve 1 has been used in the simulation shown in Fig. 2 (curve 2). The three curves in this figure are intentionally numbered 2 to 4 in order to keep the indexing of the curves the same as the corresponding ones in Fig. 2.

linearly related to the pressure drop across it, and the resonator is a linear harmonic oscillator with an amplitude-independent resonant frequency that is set by the stiffness of the diaphragm, the dimensions of the flow path, and ρ_s . When the system enters the Josephson regime, the weak link is expected to act as a nonlinear inductive element because the current through it is given by the Josephson current, which is a nonlinear function of the quantum phase. Thus, through the a.c. Josephson relation, the current depends nonlinearly on the pressure drop across the flow path. The behaviour of the resonator can now be modelled as an oscillator with an extra nonlinear contribution that is due to the current in the weak link and is given by:

$$\ddot{\phi} = -\omega_0^2[(\phi - \kappa) + LI(\phi)] = -L\omega_0^2\left[\frac{\phi - \kappa}{L} + I(\phi)\right] \quad (1)$$

Here, ω_0 is the frequency of the resonator without the weak link (with the parallel flow path only). L is the measure of the inertia of the fluid in the parallel flow path (its hydrodynamic inductance), defined such that the current through it is $(\phi - \kappa)/L$, and κ is the circulation around the loop formed by the weak link and the parallel path. From equation (1), it is clear that the link behaves as a nonlinear inductor of value $L_1 = \phi/I(\phi)$. This value changes with the amplitude of oscillations, and hence the resonance frequency of the oscillator becomes amplitude-dependent. Positive values of L_1 lead to resonant frequencies higher than ω_0 , while negative values bring it below ω_0 . This latter case can only occur in or near the ideal Josephson regime.

The current through the weak link is limited by the critical value, whereas the current through the parallel path is essentially unlimited. Hence, at a sufficiently large amplitude practically all of the current flows through the parallel path, and the resonant frequency converges to ω_0 . The shape of the amplitude dependence of the resonance frequency provides information on the $I(\phi)$ of the weak link.

The changes in the resonance frequency of the oscillator can be measured by applying variable-amplitude excitation at fixed frequency slightly off the resonance, as in ref. 1. In this case, changes of the resonant frequency affect the amplitude of the response. We used a different, more direct, technique in which the resonator was driven at resonance and the variations of the resonance frequency were recorded, as the amplitude of the response was changed via the

drive level. To a first approximation, these variations depend only on the inertia of the fluid in the link and do not depend on dissipation. On the other hand, the amplitude of the response, to a first approximation, depends only on the excitation level and the dissipation in the oscillator.

The first set of data we present (Fig. 1) shows the oscillator amplitude as a function of drive level, $A(d)$. At temperatures higher than approximately 10 mK below the transition, the oscillator amplitude increases in steps (that is, not continuously) as the drive level is increased. Such staircase patterns are similar to those exhibited by RF-SQUIDS⁹ and their superfluid analogues^{1,10} and are evidence of a periodic and hysteretic current–phase relationship. The steps are due to quantized dissipation associated with discrete 2π vortex production and show that the system is in the phase-slip regime.

Curves 1–3 in Fig. 1 show $A(d)$ for temperatures successively closer to the superfluid transition. As discussed earlier, entering the Josephson regime is expected when the critical velocity corresponds to a critical phase difference less than π : in other words, when the first step becomes approximately half the size of the subsequent 2π steps. Examining the curves in Fig. 1, we clearly see that this criterion is almost met at about 154 μ K below the transition (curve 2). At 61 μ K, the response curve has evolved from a staircase pattern into almost a straight line. Because the amplitude is sensitive primarily to dissipation, this indicates that the dissipation due to the phase slips has disappeared. Thus, the staircase data demonstrate qualitatively that the Josephson regime is being approached, and is then reached, as the temperature is brought closer to the superfluid transition. For our experiments, this occurs at $T \geq T_\lambda - 100 \mu$ K. We note that at this temperature the coherence length becomes comparable to the width of the slit-apertures constituting the weak link.

A quasiperiodic pattern is quite clear in the dependence of the resonant frequency on the amplitude of the oscillations ($f(A)$), shown in Fig. 2. This pattern occurs as a result of averaging of the nonlinear hydrodynamic inductance of the link L_l , which is periodic, over more and more periods, as the amplitude of the oscillation increases.

Most importantly, in this temperature range the frequency starts to drop to values below ω_0 (given by large-amplitude resonant frequency). This is definite proof of negative inertia, which, as it was shown earlier, can only occur in or near the ideal Josephson regime. We also note that for curve 1 the resonance frequency is amplitude-independent for small amplitudes, whereas for curve 4 it is amplitude-dependent at all drive levels. This is indicative of the substantial linear part of $I(\phi)$ in curve 1, which is absent in curve 4, taken closer to T_λ .

To summarize our observations, at 3.72 mK below T_λ , curves 1 (in both Fig. 1 and Fig. 2) show that we are deep in the phase-slip regime as shown by the amplitude-independent part and the positive (relative to ω_0) frequency shifts in $f(A)$ as well as by the steps in $A(d)$. At $T_\lambda - T = 154 \mu$ K, the amplitude-independent part of $f(A)$ is shortened (curve 2 in Fig. 2) and it reaches values below ω_0 . However, phase-slip dissipation revealed by the steps on $A(d)$ (curve 2 in Fig. 1) shows that we are not in the ideal Josephson regime yet. Finally, at $T_\lambda - T = 61 \mu$ K (curves 3 in both Figs 1 and 2) and at $T_\lambda - T = 28 \mu$ K (curve 4 in Fig. 2), the system is in ideal Josephson regime: no significant linear part in $I(\phi)$; negative frequency shifts arising from the negative inertia; and no phase-slip dissipation. This transition occurs when the critical velocity corresponds to a phase difference of the order π .

We note here that the observation of a well-defined staircase far from the Josephson regime came quite unexpectedly. Phase slips were not observed with an array of holes at lower temperatures (Y.M., A. Loshak, K. Schwab, J. C. Davis and R. E. Packard, unpublished measurements). In addition, previous results^{11–15} have suggested that fluctuations-assisted stochastic vortex creation

leads to variations of the critical velocity which should smooth out the steps entirely. In principle, it is possible that such is indeed the case in our experiments above $(T_\lambda - T) \approx 100 \mu$ K. However, this would not explain the frequency versus amplitude data. In the Josephson regime, the absence of stochastic vortex creation can be ascribed to the fact that the core of the vortex is bigger than the link size and, thus, vortex creation is suppressed. Why the fluctuations in the order parameter do not have an adverse effect on the observation of the Josephson effect remains an open question.

To lend further credence to our claim of observing the Josephson effect, we numerically simulated the effect of a weak link on the resonance frequency by solving equation (1) with dissipation, phase slips and drive terms included. The simulations support our conclusions that smooth amplitude–drive curves and smooth frequency–amplitude curves with the frequency oscillating around the large-drive frequency are observed only with continuous $I(\phi)$.

For data in the ideal Josephson regime, we extract an approximation to $I(\phi)$ directly from the curves in Fig. 2 using equation (1). To do this, we note that for small nonlinearities, the equation of motion of the cell can be solved asymptotically by assuming $\phi(t) = \Gamma \sin(\omega(\Gamma)t) + \phi_0(\Gamma)$. We then replace $I(\phi)$ by its Fourier sine expansion $I(\phi) = \sum_{k=1, \dots, \infty} I_k \sin(k\phi)$. The resonance frequency of the oscillator is then given by:

$$\left(\frac{\omega(\Gamma)}{\omega_0(\Gamma)}\right)^2 = 1 + \frac{2L}{\Gamma} \sum_{k=1, \dots, \infty} I_k J_1(k\Gamma) \cos(k\phi_0(\Gamma)) \quad (2)$$

where J_1 is a Bessel function. $\Gamma = A/(\zeta\omega)$ is the amplitude of the phase oscillations, A is the amplitude of the diaphragm motion and ζ is a characteristic of the resonator. Further details of the above model will be described elsewhere.

We fit the data to equation (2). This procedure allows us to recover any single-valued $I(\phi)$, unlike fitting to a model function, used in ref. 1. Examples of the fits are shown in Fig. 2. Once the values for the parameters I_k are obtained from the fits, the current phase characteristic is reconstructed. Close to T_λ , the current–phase relationship is found to be almost sinusoidal, as seen in Fig. 3 (curve 4).

The results we present here open up an area that was considered to be out of the reach of experimentalists. Many interesting modifications to our experiment are possible, to study effects of fluctuations on the Josephson effect, proximity effects in superfluid–normal–superfluid junctions, and applications of ⁴He weak links to the development of sensitive gyroscopes. □

Received 29 December 2000; accepted 19 March 2001.

- Avenel, O. & Varoquaux, E. Josephson effect and phase slippage in superfluids. *Jpn J. Appl. Phys.* (Suppl. 3) **26**, 1798–1803 (1987).
- Josephson, B. D. Possible new effects in superconductive tunneling. *Phys. Lett.* **1**, 251–253 (1962).
- Anderson, P. W. Consideration on the flow of superfluid helium. *Rev. Mod. Phys.* **38**, 298–310 (1966).
- Likharev, K. K. Superconducting weak links. *Rev. Mod. Phys.* **51**, 101–159 (1979).
- Backhaus, S., Pereverz, S. V., Loshak, A., Davis, J. C. & Packard, R. E. Direct measurement of the current–phase relation of a superfluid ³He–B weak link. *Science* **278**, 1435 (1997).
- Borovik-Romanov, A. S. et al. Analogy of the Josephson effect in spin supercurrent. *JETP Lett.* **47**, 478–482 (1988).
- Lifshitz, E. M. & Pitaevskii, L. P. *Statistical Physics, Part 2* Ch. 28 109–111 (Pergamon, Oxford, 1980).
- Feynmann, R. P. *Feynman Lectures on Physics* Vol. 3 Ch. 21 (McGraw-Hill, New York, 1975).
- Tinkham, M. *Introduction to Superconductivity* Ch. 6–6.5 (McGraw-Hill, New York, 1975).
- Beecken, B. P. & Zimmerman, W. Jr Search for an ac Josephson effect in superfluid ⁴He using a low frequency acoustic resonator. *Phys. Rev. B* **35**, 74–88 (1987).
- Schwab, K., Bruckner, N. & Packard, R. The superfluid ⁴He analog of the RF SQUID. *J. Low-Temp. Phys.* **110**, 1043–1104 (1998).
- Varoquaux, E., Zimmerman, W. Jr & Avenel, O. in *Proc. NATO Adv. Res. Workshop* 343–356 (Plenum, New York, 1991).
- Clow, J. & Reppey, D. Temperature dependence of superfluid critical velocities near T_λ . *Phys. Rev. Lett.* **19**, 291–293 (1967).
- Avenel, O., Aarts, R., Ihas, G. G. & Varoquaux, E. Numerical simulations of the operation of a superfluid gyrometer. *Physica B* **194–196**, 491–492 (1994).
- Day, P. et al. The fluctuation-imposed limit for temperature measurement. *J. Low-Temp. Phys.* **107**, 359–370 (1997).

Acknowledgements

We would like to thank O. Avenel, N. Bruckner, D. Goodstein, W. Holmes, K. Schwab, E. Varoquaux and P. Welander for discussions and support. This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Y.M. was funded by CEA-Saclay, France.

Correspondence and requests for materials should be addressed to Y.M. (e-mail: muh@dre.cam.saclay.cea.fr).

Conversion of silicon carbide to crystalline diamond-structured carbon at ambient pressure

Yury Gogotsi^{*,†}, Sascha Welz[†], Daniel A. Ersoy[†] & Michael J. McNallan[†]

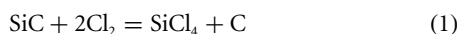
^{*} Drexel University, Department of Materials Engineering, Philadelphia, Pennsylvania 19104, USA

[†] University of Illinois at Chicago, Department of Civil and Material Engineering, 842 West Taylor Street, Chicago, Illinois 60607, USA

Synthetic diamond is formed commercially using high-pressure¹, chemical-vapour-deposition² and shock-wave³ processes, but these approaches have serious limitations owing to low production volumes and high costs. Recently suggested alternative methods of diamond growth include plasma activation⁴, high pressures⁵, exotic precursors^{6,7} or explosive mixtures⁸, but they suffer from very low yield and are intrinsically limited to small volumes or thin films. Here we report the synthesis of nano- and micro-crystalline diamond-structured carbon, with cubic and hexagonal structure, by extracting silicon from silicon carbide in chlorine-containing gases at ambient pressure and temperatures not exceeding 1,000 °C. The presence of hydrogen in the gas mixture leads to a stable conversion of silicon carbide to diamond-structured carbon with an average crystallite size ranging from 5 to 10 nanometres. The linear reaction kinetics allows transformation to any depth, so that the whole silicon carbide sample can be converted to carbon. Nanocrystalline coatings of diamond-structured carbon produced by this route show promising mechanical properties, with hardness values in excess of 50 GPa and Young's moduli up to 800 GPa. Our approach should be applicable to large-scale production of crystalline diamond-structured carbon.

We have previously shown that selective etching of carbides is an attractive technique for synthesis of carbon coatings⁹. Supercritical water¹⁰ or halogens¹¹ can be used to remove silicon from silicon carbide (SiC), producing carbide-derived carbon (CDC) films and powders that may have a variety of structures depending on the experimental conditions¹². Hydrothermal etching of SiC produced *sp*³-bonded carbon and some diamond^{10,13}, but the process was plagued by a low yield and poor reproducibility, because of the difficulty in controlling the concentrations/activities of species in high-pressure autoclaves.

Our synthesis of CDC was conducted at ambient pressure in a quartz tube furnace at or below 1,000 °C, as described in the Methods section. The SiC samples were exposed to flowing gas mixtures of 1–3.5% Cl₂, 0–2% H₂ and balance Ar, which was used as a carrier gas. Because SiCl₄ is much more thermodynamically stable than CCl₄, chlorine reacts selectively with the silicon at SiC surfaces by the reaction



thereby leaving carbon on the SiC substrate¹¹. CDC coatings

produced by treatment of sintered SiC in Ar/3.5% Cl₂ were black and graphitic, according to X-ray diffraction (XRD) and Raman spectroscopy. They showed a low hardness and Young's modulus (Fig. 1). However, cross-sectional hardness measurements using a nano-indenter showed the existence of an intermediate layer, several micrometres thick, between SiC and graphitic carbon. The material in this layer had an average hardness of about 20–30 GPa and Young's moduli of 200–300 GPa. Transmission electron microscopy (TEM) studies of this layer (Fig. 2) showed that it consisted of a mixture of graphitic carbon (as 'onions', ribbons and disordered carbon) and nanocrystalline diamond-structured carbon (Fig. 2a). In some regions, large nanocrystalline areas producing characteristic fringing (Fig. 2b) or containing microcrystals of diamond-structured carbon (Fig. 3) were found. Lattice fringing, selected-area electron diffraction (SAED), convergent-beam electron diffraction (CBED) and electron energy-loss spectroscopy (EELS) techniques confirmed the formation of diamond-structured carbon in this layer. However, lattice fringes and diffraction spots at ~0.193 nm and ~0.218 nm (Table 1), as well as characteristic CBED images (see Supplementary Information), suggest formation of 2H hexagonal diamond (lonsdaleite) along with cubic diamond-structured carbon. Lonsdaleite has often been observed in nanocrystalline diamond films^{14,15} and accompanies SiC in some natural samples¹⁶.

Addition of hydrogen to the gas to achieve a chlorine/hydrogen ratio of 2:0.575 to 2:1 resulted in changes in the appearance and structure of the carbon coatings. The layers produced at high hydrogen contents were grey, translucent in thin sections, and

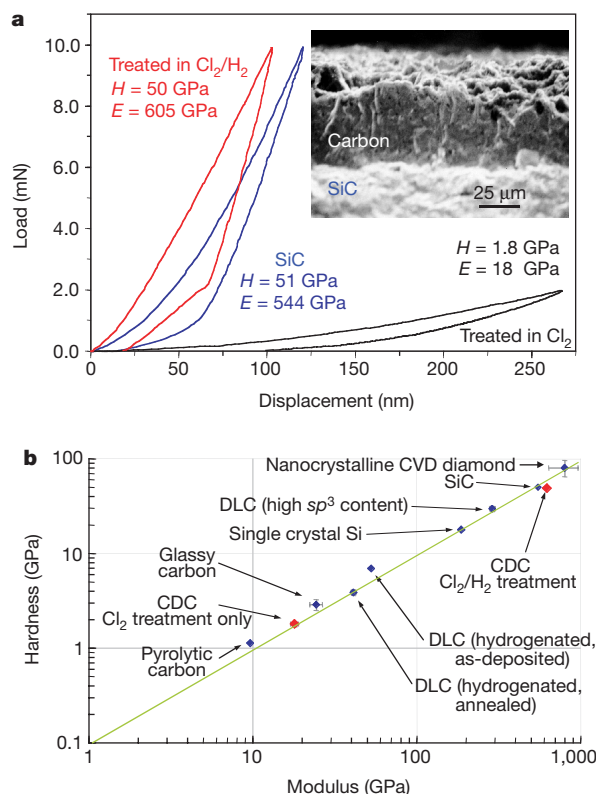


Figure 1 Results of nano-indentation tests. **a**, Typical indentation load–displacement curves obtained on polished surfaces of sintered SiC (blue curve), as well as carbon coatings produced in Ar/3.5% Cl₂ (black curve) and in Ar/2.77% Cl₂/1.04% H₂ (red curve). Inset, SEM micrograph of the hard coating produced by treatment in Ar/2.77% Cl₂/1.04% H₂ for 30 h at 1,000 °C. *H*, hardness; *E*, Young's modulus measured on the polished surface using depth-sensing indentation. **b**, Hardness versus Young's modulus for various carbon materials, silicon and SiC. Bars show standard deviations. DLC, diamond-like carbon.

fracture surfaces showed a continuous fracture pattern from the coating into SiC (Fig. 4a), suggesting that the mechanical properties of these coatings are close to that of SiC. These coatings were grown to a thickness of 50 μm (Fig. 1). CDC coatings produced with a chlorine to hydrogen ratio of about 2:1 had hardness in excess of 50 GPa and a Young's modulus of $\sim 600\text{--}800$ GPa (Fig. 1). These values exceed that of diamond-like carbon, but are below that of single-crystal or chemical vapour deposition (CVD) diamond, when measured using the same instrument (Fig. 1b). TEM showed that these coatings were built of nanocrystals, with an average size of 5–10 nm (Fig. 4b). In the SAED pattern from this film (see Supplementary Information), sharp Bragg reflections are visible up to the order of (800), indicating good crystallinity. No scattering intensity from either graphite or amorphous carbon can be seen. However, kinematically forbidden diamond reflections (200), (222) and (420) were consistently observed during this study (Table 1, Fig. 3d). These are common for Si and diamond crystals, and may appear because the allowed {111} beam acts like a new incident beam and can be rediffracted by (111) planes exciting weak (200) and (222) reflections. However, the intensity of these reflections is expected to be very low for the $Fd3m$ cubic diamond structure.

Incorporation of impurity atoms in diamond and the formation of an ordered superstructure could also cause the above-mentioned reflections. Traces of oxygen, silicon and chlorine were observed in energy-dispersive spectroscopy (EDS), but it is difficult to say if any of these elements were dissolved in the diamond lattice. Even relatively large microcrystals of diamond-structured carbon (Fig. 3a), which produced EELS spectra typical of diamond (Fig. 3c), showed unusually strong forbidden reflections—(200) and (222)—in SAED patterns (Fig. 3d). They may suggest formation of diamond polytypes¹⁴ similar to that of SiC. For example, if the symmetry is lowered from $Fd3m$ to $F43m$ (the same as in β -SiC), a high intensity of the (200) and (222) reflections is expected¹⁷. Detailed analysis of the structure of CDC will be published elsewhere.

Thus, treatment of SiC in chlorine or chlorine-hydrogen mixtures at Cl_2/H_2 ratios equal or larger than 2:1 at 1,000 °C results in the conversion of the silicon carbide to crystalline diamond-structured carbon. An increase in the hydrogen content of the gas, resulting in a 1:1 ratio and leading to the formation of HCl, resulted in very thin films or no coating at all. This may be due to a lower thermodynamic probability of the reaction

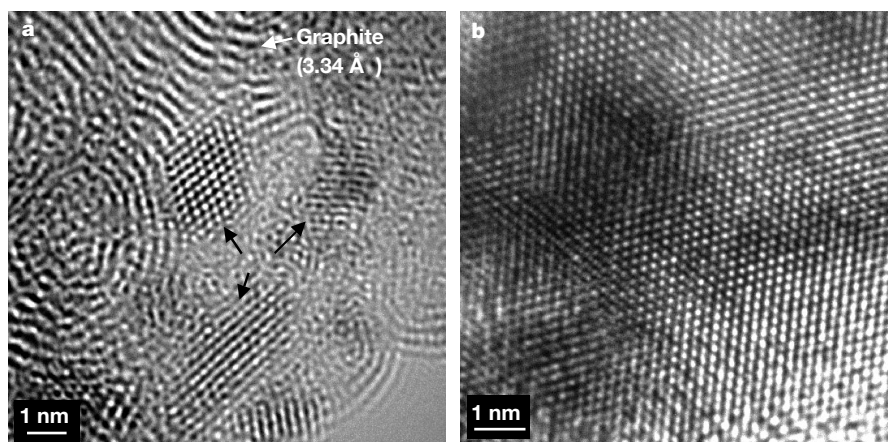
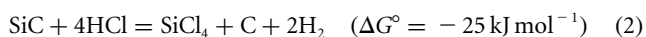


Figure 2 High-resolution TEM micrographs showing the structure of the carbon coating within a micrometre of the SiC/carbon interface. **a**, Nanocrystals of diamond-structured carbon (indicated by black arrows) surrounded by graphitic carbon (shown by white

Table 1 Comparison of d spacings

Experimental*	Cubic diamond $Fd3m$ (JCPDS 6-0675)		Lonsdaleite $P63/mmc$ (JCPDS 19-0268)	
	hkl	d spacing (nm)	hkl	d spacing (nm)
0.218–0.222			100	0.218
0.206–0.207	111	0.206	002	0.206
0.192–0.194			101	0.193
0.178–0.182	200†	0.178		
0.150			102	0.150
0.126	220	0.126	110	0.126
0.117			103	0.116
0.110			020	0.1092
0.106	311	0.1075	112	0.1075

*Typical values from SAED, CBED and lattice fringe measurements on about 30 diamond single crystals (5–500 nm in size) from three different samples.

†Forbidden cubic diamond reflection, typical for n -diamond²².

compared to reaction (1), which has $\Delta G^\circ = -434.1 \text{ kJ mol}^{-1}$ (ref. 11) where ΔG° is Gibbs free energy change.

In Ar- Cl_2 and Ar- Cl_2/H_2 environments, the thickness of the carbon layer increases linearly with time, following the equation $d = k_1 t$, where d is the layer thickness, t is time, and k_1 is the linear rate constant. For the Ar/2.77% Cl_2 /1.04% H_2 environment, $k_1 = 1.6 \mu\text{m h}^{-1}$, which is only about 20% lower than for Ar/3.5% Cl_2 . Because the kinetics are linear, the controlling factor of the reaction is not the diffusion of reactant species through the growing carbon layer. If this were the case, we would expect a parabolic rate equation¹⁸. In order for the chlorination reaction to proceed, two molecules of $\text{Cl}_2(\text{g})$ or four molecules of $\text{HCl}(\text{g})$ must be transported to the SiC/C interface, and one molecule of $\text{SiCl}_4(\text{g})$ —as well as two molecules of $\text{H}_2(\text{g})$ —must be transported away from the interface for each atom of carbon produced. Linear kinetics implies that the carbon film is nanoporous and allows easy permeation of Cl_2 , HCl , H_2 and SiCl_4 molecules, in spite of its dense appearance in scanning electron microscopy (SEM; Fig. 4a) at magnifications up to $\times 500,000$. This nanoporosity may be responsible for the hardness values being lower than that of CVD diamond (Fig. 1b). An important implication of the linear layer-growth kinetics is the possibility of growing very thick coatings on SiC, or the complete conversion of SiC powders or components into diamond-structured carbon.

Recent studies have shown that it is not difficult to form nanometre-sized diamonds¹⁹. Moreover, several groups have reported the nucleation of sp^3 -bonded carbon and nanocrystalline diamond after surface treatment of SiC by fluorocarbon plasma²⁰

arrows), including onion-like structures. **b**, Region of nanocrystalline diamond-structured carbon. The sample was treated for 24 h at 1,000 °C in Ar/3.5% Cl_2 .

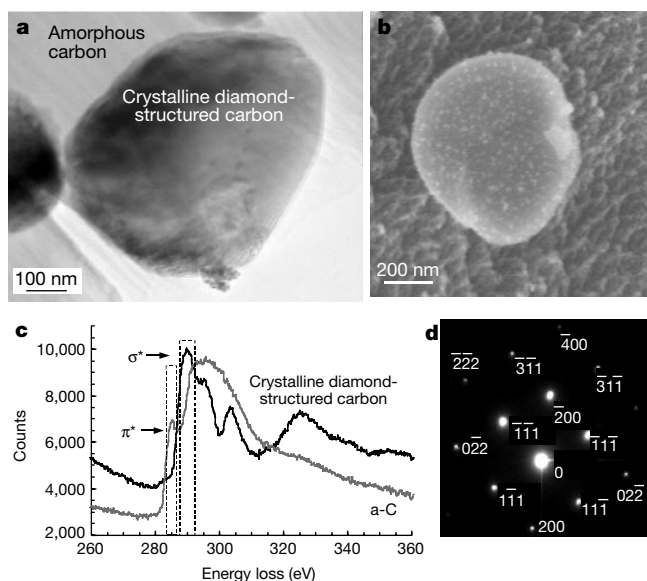


Figure 3 Characterization of crystalline diamond-structured carbon embedded in amorphous carbon. **a**, TEM image. **b**, SEM image. **c**, EELS spectra from the particle and surrounding carbon material in **a**. **d**, [011] SAED pattern from the particle imaged in **a**. SAED and the EELS spectra show that the rounded particles are diamond-structured carbon. The light material in **a** was identified as carbon with a high content of sp^3 bonding and some sp^2 bonding using EELS (**c**). Etching in hydrogen plasma selectively removes amorphous and disordered carbon, revealing crystals of diamond-structured carbon in the carbon layer (**b**). The rounded shape of the microcrystals can be explained by their solid-state growth via coalescence of nanocrystals when constrained by the carbon matrix.

and bombardment with hydrogen²¹ or carbon ions²². Thus, there is consistent evidence of conversion of carbides into diamond, after removal of metal atoms from the carbide lattice under various experimental conditions. The growth of diamond-structured carbon from SiC in pure chlorine with no hydrogen added (Figs 2 and 3) is also in agreement with the synthesis of diamond from fullerene^{6,19}, which showed that hydrogen is not essential for diamond growth outside its range of thermodynamic stability. According to the original concept for the metastable growth of diamond suggested by Spitsyn²³, it is necessary to conserve the orientational effect of the surface carbon atoms and to use carbon-containing molecules with sp^3 bonding that can be attached to the

diamond surface in a complementary manner. Both conditions can be satisfied when Si is extracted from SiC, forming carbon atoms in sp^3 hybridization. We note that the first diamond growth experiments of Derjagin and Spitsyn²³ were conducted in the carbon-halogen system, using CBr_4 and Cl_4 . It can be assumed that the tetrahedrally coordinated SiC lattice, which is preserved during chlorination, acts as a template for the growth of diamond^{10,11}—and that diamond-structured carbon grows by direct transformation of the SiC lattice because of the sp^3 bonding of carbon in SiC. The work we report here was performed using α -SiC; but we note that β -SiC has a similar structure, namely a diamond lattice where 50% of carbon atoms are replaced with Si, and that this polytype could also be converted to crystalline diamond-structured carbon (results available, but not shown here). We have therefore shown that both SiC polytypes can be converted to crystalline diamond-structured carbon, although the question remains as to whether the carbide lattice structure (cubic or hexagonal) affects the structure of the resulting CDC. Molecular-dynamics simulations using empirical interatomic Tersoff potentials have shown that for a Si-terminated (1000) 6H-SiC surface, very high lattice strains do not allow direct growth of diamond on SiC, and fragmentation—leading to nanocrystalline material—must occur²⁴. Small diamond clusters on SiC are predicted to have a good adhesion to the substrate and maintain sp^3 coordination of carbon atoms in the cluster²⁴. Our TEM study of the chlorine-treated SiC suggested that SiC was converted to amorphous sp^3 carbon, and formation of crystalline diamond-structured carbon occurred from disordered, Si-depleted SiC within nanometres of the SiC/carbon interface. Random orientation of nanocrystals of diamond-structured carbon (Fig. 2) in CDC supports non-epitaxial growth of diamond-structured carbon. Thus, growth of nanocrystalline diamond-structured carbon occurred from highly disordered sp^3 carbon produced by selective etching of SiC. Growth of larger crystals of diamond-structured carbon (Fig. 3a, b) was probably the result of coalescence of continuous nanocrystalline regions (Fig. 2b). However, if no hydrogen was added to the gas, nanocrystalline diamond-structured carbon was slowly transformed to the thermodynamically stable (under ambient pressure) graphitic carbon during the long-term treatment at 1,000 °C, and we observed only amorphous and graphitic carbon at a distance of more than 3 μm from the SiC/carbon interface. Thus, the role of hydrogen is primarily in stabilization of dangling bonds of carbon. This helps to maintain sp^3 hybridization of carbon and prevent formation of sp^2 -bonded carbon. Therefore, addition of hydrogen stabilized the diamond-structured carbon phase and allowed the continuous growth of a film of this phase on the surface. Theoretical aspects of hydrogen-assisted diamond growth at low pressures²³ and the thermodynamic

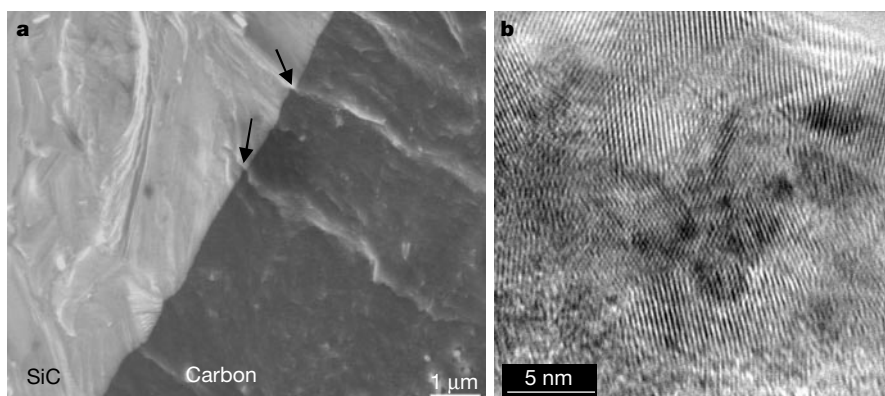


Figure 4 Microstructure of a carbon film produced in Cl_2/H_2 . **a**, SEM micrograph of a fracture surface showing a carbon layer over the SiC substrate. **b**, Typical TEM micrograph of the nanocrystalline layer of diamond-structured carbon produced with

hydrogen present in the reaction gas. Sample was sintered α -SiC treated in $Ar/2.77\% Cl_2/1.04\% H_2$ for 30 h at 1,000 °C. Arrows in **a** show a continuous fracture pattern from the coating into SiC.

coupling model that can be used to explain the formation of diamond in reactions (1) and (2) have been well developed²⁵, and will not be discussed here.

The transformation of SiC to crystalline diamond-structured carbon that we observed can explain diamond-SiC paragenesis²⁶, and probably the formation of carbonado diamonds in nature²⁷. It also shows that the occurrence of diamond/lonsdaleite in association with SiC can be caused by reasons other than hypervelocity impact on Earth¹⁶.

The manufacture of CDC has been developed and scaled up during the past few years^{12,28}, and a variety of useful products—ranging from nanoporous carbons²⁹ for supercapacitors and batteries, to nanotubes, onion-like carbon³⁰, and tribological coatings³¹—have been reported. Some of these have been commercialized or are on their way to commercialization. Facilities for halogenation of carbides, which are able to produce tons of carbon per day, have been created (for example, see <http://www.skeleton-technologies.com>). Only minor modification of the process (hydrogen addition) is required for synthesis of crystalline diamond-structured carbon using the same equipment. The process that we have developed is versatile because it allows synthesis of powders or coatings of virtually any thickness. As the transformation is conformal and does not change the shape of the particle, powders with any grain size could be produced by using raw SiC powders of different particle sizes. However, the product will be grains built of nanocrystals of diamond-structured carbon. The presence of micrometre-size crystals in our samples (Fig. 3a) shows that coalescence of nanocrystals of diamond-structured carbon may occur under appropriate conditions, and lead to growth of microcrystalline diamond-structured carbon.

The mechanical, electrical and optical properties of nanocrystalline diamond-structured carbon are altered from those of single diamond crystals, because the grain-boundary carbon is π -bonded, as shown by the Raman spectra and EELS. The fraction of atoms residing at grain boundaries can be up to 10% when the average crystallite size is 3–5 nm (ref. 19). Nanocrystalline diamond films have found applications in tribology: for example, to coat SiC dynamic seals for water pumps¹⁹. Graphitic carbon in CDC films can act as a solid lubricant³¹. Remarkable electron emission with turn-on fields of $\sim 1 \text{ V } \mu\text{m}^{-1}$ and a high current density has been achieved using thin films of nanocrystalline diamond¹⁹. Nanocrystalline diamond has a higher conductivity than boron-doped microcrystalline diamond, and can be used for electrodes in chemically aggressive environments¹⁹. Conformal coatings produced by selective etching¹¹ would be useful in microelectromechanical system (MEMS) applications where very thin and uniform coatings are required. In addition, permeability of the films produced by chlorination of SiC and an extremely narrow pore-size distribution in CDC²⁹ may enable the production of molecular sieves, high-surface-area electrodes and other applications, where vapour-deposited diamond films cannot be applied. The large-scale, solid-state synthesis that we report here, being relatively low-cost, and operating at ambient pressure and moderate temperatures with no plasma activation, may open the way for diamond-structured materials to be used in a variety of high-volume applications—such as brake pads. □

Methods

Synthesis

The apparatus used to investigate the interactions of SiC with halogens at atmospheric pressure has been described elsewhere¹¹. Experiments were performed using several commercially available β -SiC powders, sintered α -SiC and CVD β -SiC materials; however, the experiments described here are limited to sintered α -SiC. These samples were sectioned into disks 16 mm in diameter and 1 mm thick. The disks were cleaned ultrasonically, rinsed in acetone, and placed in a quartz sample holder. This was in turn suspended via a silica wire connected to a fused-silica rod in the centre of a fused-silica reaction tube in the hot zone of a furnace. Experiments were continued for between 30 min and 30 h over a broad temperature range (600–1,100 °C). The results presented here were

obtained in experiments conducted at 1,000 °C. At the end of each experimental run, an argon purge was initiated through the reaction chamber during the cool-down period.

Analysis

The reacted specimens were analysed using optical microscopy and SEM, EDS, EELS, XRD and Raman spectroscopy (Ar-ion laser, 514.5-nm excitation wavelength). JEOL JEM-3010 (300 kV) and JEOL JEM-2010F (200 kV) TEMs, as well as a JEOL 6320 field-emission SEM, were used for this work. EDS was used to identify the carbon areas which were free from impurities and showed only traces of silicon and chlorine, and EELS was used to identify areas of diamond-structured carbon from the carbon edge shape. Subsequently, SAED was performed on $\sim 1\text{-}\mu\text{m}$ nanocrystalline areas and microcrystals, and CBED was performed with a 5-nm and 10-nm electron probe. A Nano Indenter XP (MTS) equipped with a Berkovich indenter (diamond pyramid) was used to measure the hardness and Young's modulus of the coating under a load of 10 mN and a loading rate of $100 \mu\text{N s}^{-1}$.

Received 5 December 2000; accepted 9 February 2001.

- Bundy, F. P., Hall, H. T., Strong, H. M. & Wentorf, R. H. Man-made diamonds. *Nature* **176**, 51 (1955).
- Angus, J. C. & Hayman, C. C. Low-pressure, metastable growth of diamond and "diamondlike" phases. *Science* **241**, 913–921 (1988).
- Burkhard, G., Dan, K., Tanabe, Y., Sawaoka, A. B. & Yamada, K. Carbon phase transition by dynamic shock compression of a copper/graphite powder mixture. *Jpn J. Appl. Phys.* **33**, L876–L879 (1994).
- Roy, R. *et al.* New process for 1 atm diamond synthesis: from metallic solutions. *Innov. Mater. Res.* **1**, 65–87 (1996).
- Zhao, X.-Z., Roy, R., Chierian, K. A. & Badzian, A. Hydrothermal growth of diamond in metal-C-H₂O systems. *Nature* **385**, 513–515 (1997).
- Gruen, D. M., Liu, S., Krauss, A. R. & Pan, X. Buckyball microwave plasmas: fragmentation and diamond-film growth. *J. Appl. Phys.* **75**, 1758–1763 (1994).
- Regueiro, M. N., Monceau, P. & Hodeau, J.-L. Crushing C₆₀ to diamond at room temperature. *Nature* **355**, 237–239 (1992).
- Li, Y. *et al.* A reduction-pyrolysis-catalysis synthesis of diamond. *Science* **281**, 246–247 (1998).
- Gogotsi, Y. G. & Yoshimura, M. Formation of carbon films on carbides under hydrothermal conditions. *Nature* **367**, 628–630 (1994).
- Gogotsi, Y. G., Kofstad, P., Nickel, K. G. & Yoshimura, M. Formation of sp^3 -bonded carbon upon hydrothermal treatment of SiC. *Diamond Relat. Mater.* **5**, 151–162 (1996).
- Gogotsi, Y. G., Jeon, J. D. & McNallan, M. J. Carbon coatings on silicon carbide by reaction with chlorine-containing gases. *J. Mater. Chem.* **7**, 1841–1848 (1997).
- Gogotsi, Y. in *Proc. NATO ARW on Nanostructured Films and Coatings* (eds Chow, G.-M., Ovid'ko, I. A. & Tsakalakos, T.) 25–40 (Kluwer, Dordrecht, 1999).
- Gogotsi, Y. G., Nickel, K. G. & Kofstad, P. Hydrothermal synthesis of diamond from diamond-seeded β -SiC powder. *J. Mater. Chem.* **5**, 2313–2314 (1995).
- Silva, S. R. P., Amarantunga, G. A. J., Salje, E. K. H. & Knowles, K. H. Evidence of hexagonal diamond in plasma-deposited carbon films. *J. Mater. Sci.* **29**, 4962–4963 (1994).
- Zarrabian, M., Fourches-Coulon, N., Turban, G., Marhic, C. & Lancin, M. Observation of nanocrystalline diamond in diamondlike carbon films deposited at room temperature in electron cyclotron resonance plasma. *Appl. Phys. Lett.* **70**, 253–255 (1997).
- Hough, R. M. *et al.* Diamond and silicon carbide in impact melt rock from the Ries impact crater. *Nature* **378**, 41–44 (1995).
- Rossi, M., Vitali, G., Terranova, M. L. & Sessa, V. Experimental evidence of different crystalline forms in chemical vapour deposited diamond films. *Appl. Phys. Lett.* **63**, 2765–2767 (1993).
- Kofstad, P. *High Temperature Corrosion* (Elsevier, London, 1988).
- Gruen, D. M. Nanocrystalline diamond films. *Annu. Rev. Mater. Sci.* **29**, 211–259 (1999).
- Grannen, K. J. & Chang, R. P. H. Diamond growth on carbide surfaces using a selective etching technique. *J. Mater. Res.* **9**, 2154–2163 (1994).
- Lannon, J. M. J., Gold, J. S. & Stinespring, C. D. Hydrogen ion interactions with silicon carbide and the nucleation of diamond thin films. *J. Appl. Phys.* **77**, 3823–3830 (1995).
- Heera, V., Skorupa, W., Pecz, B. & Dobos, L. Ion beam synthesis of graphite and diamond in silicon carbide. *Appl. Phys. Lett.* **76**, 2847–2849 (2000).
- Spitsyn, B. V. in *Handbook of Crystal Growth* (ed. Hurler, D. T. J.) 401–456 (Elsevier, London, 1994).
- Gogotsi, Y. *et al.* in *Proc. NATO ASI on Functional Gradient Materials and Surface Layers Prepared by Fine Particles Technology* (eds Baraton, M.-I. & Uvarova, I. V.) 239–255 (Kluwer, Dordrecht, 2001).
- Wang, J.-T., Cao, C.-B. & Zheng, P.-J. Theoretical aspects of low pressure diamond synthesis. *J. Electrochem. Soc.* **141**, 278–281 (1994).
- Leung, I., Guo, W., Friedman, I. & Gleason, J. Natural occurrence of silicon carbide in a diamondiferous kimberlite from Fuxian. *Nature* **346**, 352–354 (1990).
- Daulton, T. L. & Ozima, M. Radiation-induced diamond formation in uranium-rich carbonaceous materials. *Science* **271**, 1260–1262 (1996).
- Gogotsi, Y. *et al.* Formation of carbon coatings on SiC fibres by selective etching in halogens and supercritical water. *Ceram. Eng. Sci. Proc.* **19**, 87–94 (1998).
- Kyutt, R. N., Smorgonskaya, E. A., Danishevskii, A. M., Gordeev, S. K. & Grechinskaya, A. V. Structural study of nanoporous carbon produced from polycrystalline carbide materials: small angle X-ray scattering. *Phys. Solid State* **41**, 1359–1363 (1999).
- Zheng, J., Ekström, T. C., Gordeev, S. K. & Jacob, M. Carbon with an onion-like structure obtained by chlorinating titanium carbide. *J. Mater. Chem.* **10**, 1039–1041 (2000).
- Ersoy, D. A., McNallan, M. J., Gogotsi, Y. & Erdemir, A. Tribological properties of carbon coatings produced by high temperature chlorination of silicon carbide. *STLE Tribol. Trans.* **43**, 809–815 (2000).
- Hirai, H. & Kondo, K.-I. Modified phases of diamond formed under shock compression and rapid quenching. *Science* **253**, 772–774 (1991).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank A. Nicholls for discussions and help with TEM analysis. The work done at UIC was supported by the US NSF, and the work done at Drexel University was supported by DARPA via ONR contract. The electron microscopes used in this work are operated by the Research Resources Center at UIC. The JEM-2010F purchase was supported by the NSF.

Correspondence and requests for materials should be addressed to Y.G.
(e-mail: gogotsi@drexel.edu).

A 300-million-year record of atmospheric carbon dioxide from fossil plant cuticles

Gregory J. Retallack

Department of Geological Sciences, University of Oregon, Eugene, Oregon 97403-1272, USA

To understand better the link between atmospheric CO₂ concentrations and climate over geological time, records of past CO₂ are reconstructed from geochemical proxies^{1–4}. Although these records have provided us with a broad picture of CO₂ variation throughout the Phanerozoic eon (the past 544 Myr), inconsistencies and gaps remain that still need to be resolved. Here I present a continuous 300-Myr record of stomatal abundance from fossil leaves of four genera of plants that are closely related to the present-day *Ginkgo* tree. Using the known relationship between leaf stomatal abundance and growing season CO₂ concentrations^{5,6}, I reconstruct past atmospheric CO₂ concentrations. For the past 300 Myr, only two intervals of low CO₂ (<1,000 p.p.m.v.) are inferred, both of which coincide with known ice ages in Neogene (1–8 Myr) and early Permian (275–290 Myr) times. But for most of the Mesozoic era (65–250 Myr), CO₂ levels were high (1,000–2,000 p.p.m.v.), with transient excursions to even higher CO₂ (>2,000 p.p.m.v.) concentrations. These results are consistent with some reconstructions of past CO₂ (refs 1, 2) and palaeotemperature records⁷, but suggest that CO₂ reconstructions based on carbon isotope proxies^{3,4} may be com-

promised by episodic outbursts of isotopically light methane^{8,9}. These results support the role of water vapour, methane and CO₂ in greenhouse climate warming over the past 300 Myr.

As atmospheric CO₂ levels have risen over the past 200 years of industrial fossil fuel consumption, plants have responded by decreasing the density of stomates on their leaves^{5,6}. Here I use the well-known inverse relationship between atmospheric CO₂ concentration and stomatal density as a palaeobarometer of atmospheric CO₂ during growth of fossil plant leaves. Stomatal density of living plants has, however, been shown to be related to differences in insolation, water stress and stomatal position within leaves, which also affect cell size, and thus stomatal density. The effects of these competing variables are minimized by using stomatal index (percentage stomates over stomates plus epidermal cells) rather than stomatal density (stomates per unit area⁶). Further limits to such competing variables are suggested by the habitats of the fossil leaves studied: these were mostly humid, lowland, fluvial and swamp environments suitable for preservation of fossil plant cuticles¹⁰. The plants had nutrient-poor peaty and siliceous substrates, suffered periodic flood disturbance and were in open vegetation early in the ecological succession after disturbance, as is indicated by sedimentological and taphonomic studies of several of the studied fossil sites^{10,11}. Another potentially misleading effect is the differences between trees of different sex in *Ginkgo*: pollen-producing trees

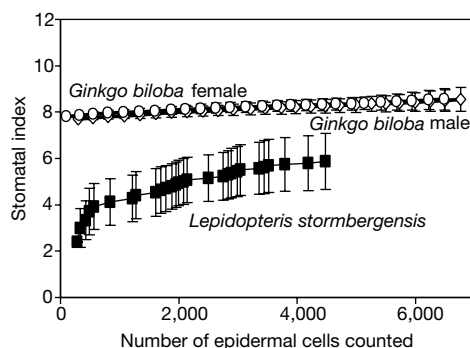


Figure 1 Rarefaction analysis indicates how many cells need to be counted for reliable determination of stomatal index (SI). Shown here are leaves of a female (circles) and male (diamonds) tree of *Ginkgo biloba* collected from Eugene, Oregon, 5 September 2000, and of *Lepidopteris stormbergensis* from the late Triassic (Carnian) Molteno formation at Little Switzerland, South Africa¹². The *Ginkgo* samples were ideal: images of the same size from different leaves of two trees. The *Lepidopteris* samples were numerous but far from ideal: images of irregular shape and varied size, with great natural variation in SI, and from an unknown number of leaves or plants.

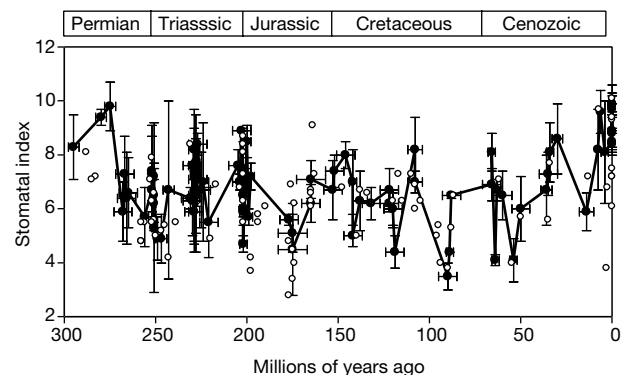


Figure 2 Stomatal index has varied considerably over the past 300 million years, but was as high during the early Permian ice age as it has been during the Neogene. Data are from fossil leaves of *Rhachiphyllum* (early Permian, 295–265 Myr ago), *Lepidopteris* (mid-Permian to late Triassic, 258–200 Myr ago), *Tatarina* (late Permian, 252–250 Myr ago) and *Ginkgo* (late Triassic to recent, 229–0 Myr ago). Filled symbols and the solid line are reliable data as established by rarefaction analysis (Fig. 1); unfilled symbols are statistically inadequate samples, and are plotted to show proven data density of the cuticular record of these taxa. All data are available as Supplementary Information.

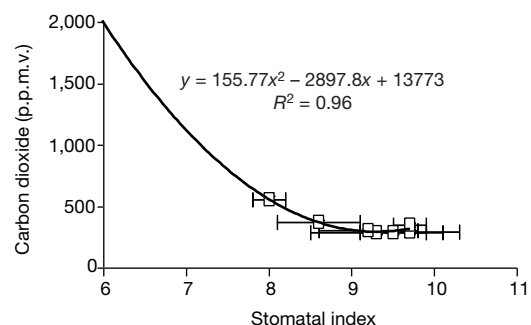


Figure 3 Palaeoatmospheric CO₂ can be inferred from fossil leaf SI by using this transfer function derived from *Ginkgo* leaves grown in greenhouse experiments⁵ and taken from herbarium sheets dating back to 1888 (see Supplementary Information).

have branches reaching upwards and so are more evenly lit, whereas horizontal branches tend to shade lower branches in ovule-producing trees. Counts of stomatal index from SEM images of leaves of a male and female tree from Eugene, Oregon, showed identical stomatal index (mean SI was 8.6 ± 0.4 ; Fig. 1).

The most serious obstacle to the use of SI as a CO₂ palaeo-barometer has been the different SI among species growing in the same location. For this reason, applications of this technique have focused on particular species, such as *Quercus petraea*, known in the fossil record back to the Miocene (10 Myr ago)⁶. Here I used primarily the genus *Ginkgo*, which has a fossil record of well-preserved leaf cuticles at least back to the late Triassic period (229 Myr ago)¹². To extend this record back into the Palaeozoic, the pteridosperm genera *Lepidopteris* (Permian–Triassic), *Tatarina* (late Permian), and *Rhachiphyllum* (Permian) were chosen for several reasons. They have similar SI where they occur in the same places: for example, *Ginkgo telemachus* (SI 7.6 ± 1.4) and *Lepidopteris africana* (SI 6.5 ± 1.7) at Little Switzerland (late Triassic, South Africa¹²), and *Tatarina conspicua* (SI 5.5 ± 1.1) and *Lepidopteris* sp. (SI 6.3 ± 0.9) at Baizovka (late Permian, Russia¹³). Furthermore, there is a phylogenetic relationship between ginkgos and peltasperm seed ferns, as revealed by analysis of their fossilized reproductive structures¹⁰. All four genera also have morphologically similar stomates. Their subsidiary cell cuticles are darker (because they are thicker) than those of other epidermal cells. Subsidiary cells that are arranged in an even ring each have a hollow papilla oriented partly to occlude the stomatal pit. This kind of stomate is also found in cycads such as *Dioon*, and conifers such as *Torreya*, *Agathis* and *Thuja*¹⁴. Such partial papillar occlusion gives

the stomatal apparatus of *Ginkgo*, *Lepidopteris*, *Tatarina*, and *Rhachiphyllum* a degree of functional equivalence that is not found in other fossil genera such as *Dicroidium*, *Pachypteris*, or *Baiera*¹².

This study used previously published illustrations of fossil cuticle (see Methods), with the exception of two early Triassic (250–245 Myr ago) species from near Sydney, Australia (see Supplementary Information). The numerical ages of the fossil cuticles were taken from a current geological time scale¹⁵, but relative age came from published accounts of the cuticles. My compilation demonstrates that SI in Permian (275–296 Myr ago) *Rhachiphyllum* and *Lepidopteris* was comparable with Neogene (0–8 Myr ago) *Ginkgo*. Such high SI (>9) reflects low levels of atmospheric CO₂, predicted for the Permian by sedimentary mass balance models¹. Cool climates would be predicted from the role of CO₂ as a greenhouse gas³, and both the Neogene and the Permian are known from evidence of tillites, striated pavements, and ice-rafted debris to be times of large global ice caps¹¹. These two end-points (Fig. 2) indicate that Permian plants were similar to modern ones in their relationship between stomatal index and atmospheric CO₂, and that there has not been long-term drift in this proxy measure of atmospheric CO₂.

Fossil plant cuticular SI can be calibrated to atmospheric CO₂ levels from greenhouse experiments using living *Ginkgo biloba*⁵, which have shown SI 9.7 ± 0.2 at 350 p.p.m.v. and SI 8.0 ± 0.2 at 560 p.p.m.v. CO₂. Herbarium specimens of *Ginkgo biloba* ranging back in age to 1888 were also counted for stomatal index (see Supplementary Information), and CO₂ levels for each taken from global estimates for their year of collection⁶. Herbarium and green-

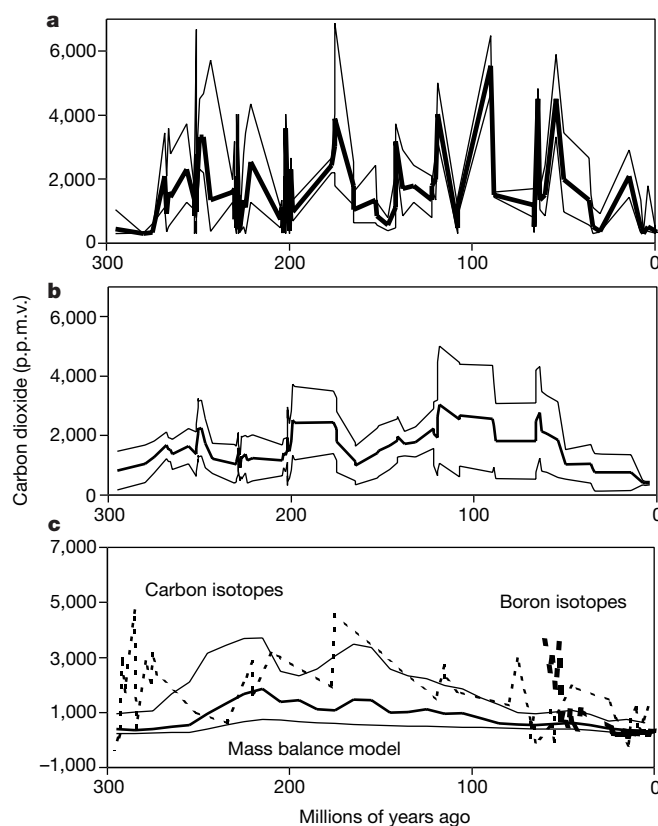


Figure 4 Cuticular estimates (**a**, **b**) of atmospheric CO₂ (p.p.m.v.) can be compared with other proxies (**c**) over the past 300 million years. **a**, Unsmoothed data using transfer function (Fig. 3) on mean and standard deviation of 84 reliable SI determinations of fossil plants (Fig. 2). **b**, Five-point moving average of transformed SI determinations with

flanking standard error of means. **c**, Previously published curves derived from a sedimentary mass balance model¹ (thick line), boron isotopic composition of marine foraminifera² (bold dashed line) and carbon isotopic composition of pedogenic carbonate³ (dashed line).

house results, together with additional greenhouse studies in progress¹⁶, indicate that the response of SI to CO₂ concentrations is markedly nonlinear. Greenhouse⁵ and herbarium data were fitted with a quadratic equation for the concentration of CO₂: [CO₂] = 155.77 (SI)² - 2897.8 (SI) + 13773, where $r^2 = 0.96$ (adjusted to sample size of 8). This transfer function (Fig. 3) was then used to estimate past partial pressures of CO₂ from individual SI measurements (Fig. 4a) and smoothed by a five-point moving average to give a curve (Fig. 4b) comparable to past estimates of atmospheric CO₂ (Fig. 4c).

Comparison of the new results with mass balance model estimates (Fig. 4c) is good, given the coarse resolution (10 Myr intervals) of the model compared with the present study (mean sample gap 3.6 ± 5.4 Myr). Cuticular estimates also show high CO₂ levels during the early Tertiary comparable to those by high-resolution (mean sample gap 1.8 ± 2.7 Myr) boron isotopic studies of marine foraminifera (Fig. 4c), although these estimates could be substantially reduced by correction for secular changes in riverine boron input to the ocean¹⁷. The increased warmth and weathering during times of high CO₂ demonstrated here is supported by evidence from petrographic and chemical studies of palaeosols^{8,11}, the migration of thermophilic organisms to high latitudes^{8,11}, and oxygen isotopic analyses of marine skeletal remains from different palaeolatitudes^{7,8}. There are also indications from palaeosols and from palynology that times of high CO₂ were times of humid palaeoclimate, and conversely that cool intervening intervals were dry. For example, the abundance of fossil pollen that is taken as evidence of aridity in many parts of Asia¹⁹ rises at all four times of low CO₂ and cooling during the Jurassic and Cretaceous (Fig. 4b). The low CO₂ and cooling from Eocene to Oligocene epochs that is indicated by the cuticular record (Fig. 4b) is matched by evidence for synchronous palaeoclimatic drying in the palaeosol record of central and western North America¹⁸.

The cuticular time series (Fig. 4b) shows numerous transient excursions to very high CO₂ (>2,000 p.p.m.v.). High-resolution studies of the SI minimum in fossil *Ginkgo* and cycad leaves across the Triassic–Jurassic boundary (200 Myr ago; ref. 20) in Greenland and Sweden indicates a transient CO₂ spike coincident with excursion to isotopically lighter carbon ($\delta^{13}C_{org}$) of the same leaves, and mass extinction (claiming *Lepidopteris* among others)²¹. Other CO₂ and carbon isotopic transients following mass extinction of the earliest Triassic (250 Myr ago), and faunal overturn of the early Jurassic period (190 Myr ago), early Cretaceous period (117 Myr ago) and late Palaeocene epoch (55 Myr ago), have been related to catastrophic outbursts of isotopically light methane from permafrost and marine hydrate reservoirs^{8,9,22}. Yet another CO₂ spike at the Cretaceous–Tertiary boundary is coincident with an excursion to lighter carbon isotopic values on land and in the sea, asteroid impact in Yucatan, and the mass extinctions that claimed dinosaurs and ammonites²³. Transient CO₂ maxima in the cuticular time series (Fig. 4a, b) represent strong perturbations of the carbon cycle.

There are significant discrepancies between cuticular estimates of atmospheric CO₂ (Fig. 4a, b) and estimates based on the carbon isotopic composition of palaeosol carbonates (Fig. 4c), which indicate CO₂ minima (including even negative concentrations) at 250 Myr ago (earliest Triassic), 190 Myr ago (early Jurassic), 117 Myr ago (early Cretaceous) and 55 Myr ago (latest Palaeocene)³. The timing of these events may be a clue to reasons for the discrepancies, because carbon isotopic studies indicate that these were times of catastrophic release of isotopically light methane from permafrost and marine gas-hydrate reservoirs^{8,9,22}. Once in the atmosphere, methane is oxidized within 2–7 years to carbon dioxide, which retains the unusual isotopic signature of gas-hydrate reservoirs isolated from global surficial systems²⁴. Atmospheric additions of isotopically light methane may also explain why the palaeosol isotopic palaeobarometer³ and high-resolution carbon isotopic studies of deep oceanic organic matter⁴ failed to detect the

high CO₂ levels and warm palaeoclimate of the middle Miocene, which is evident from the few Miocene results presented here (Figs 2, 4a, b), as well as from studies of foraminifera²⁵, plants²⁶, palaeosols²⁷ and oxygen isotopic composition of marine shells^{7,25}. The observation that carbon isotopic palaeobarometers fail to show high CO₂ at times of warm palaeoclimate has led to the suggestion that climatic warming and increasing atmospheric CO₂ concentrations were decoupled^{7,28}. If carbon isotopic CO₂ palaeobarometers are compromised by episodic methane outbursts, as suggested here, then the accepted role of CO₂, methane and water vapour as greenhouse gases remains intact for at least the past 300 million years⁷. □

Methods

Counts were made by marking photocopies of microphotographs of cuticles taken either in plain unpolarized light or from the SEM, which is preferable to light microphotographs of species with thin cuticles such as living *Ginkgo biloba*²⁹. Although this meta-analysis has the advantage of relying almost entirely on data available in the public domain, a disadvantage is that many published cuticular studies illustrate insufficient cuticle for statistically adequate determination of SI (see Supplementary Information). Rarefaction analysis was used to determine which counts were reliable. Individual determinations of SI from different illustrated cuticle fragments were arranged in order from lowest to highest and cumulative means and standard deviations were computed and plotted against numbers of epidermal cells counted. Mean SI from counting 500 epidermal cells was a good approximation of mean SI after counting 6,000 epidermal cells (Fig. 1). The accuracy of SI estimates is comparable to that of point counting the percentage mineral composition of petrographic thin sections, for which 500 points is routine, but larger counts are needed for high-precision work³⁰. Using the criterion of more than 500 epidermal cells gave 84 reliable SI measurements within the past 300 million years (not counting modern leaves). Open symbols in Fig. 2 show an additional 74 estimates based on less than 500 counted epidermal cells, and demonstrate the potential for published fossil plant cuticles to improve temporal resolution from a mean time gap between samples of 3.6 ± 5.4 Myr to 1.7 ± 2.7 Myr.

Received 13 November 2000; accepted 22 March 2001.

- Berner, R. A. The rise of plants and their effect on weathering and atmospheric CO₂. *Science* **276**, 543–546 (1997).
- Pearson, P. N. & Palmer, M. R. Atmospheric carbon dioxide concentrations over the past 60 million years. *Nature* **406**, 695–699 (2000).
- Ekart, D. P., Cerling, T. E., Montañez, I. P. & Tabor, N. J. A 400 million year carbon isotope record of pedogenic carbonate: implications for paleoatmospheric carbon dioxide. *Am. J. Sci.* **299**, 805–827 (1999).
- Pagani, M., Freeman, K. H. & Arthur, M. A. Late Miocene atmospheric CO₂ concentration and expansion of C₄ grasses. *Science* **285**, 876–879 (1999).
- Beerling, D. J., McElwain, J. C. & Osborne, C. P. Stomatal responses of the “living fossil” *Ginkgo biloba* L. To changes in atmospheric CO₂ concentrations. *J. Exp. Bot.* **49**, 1603–1607 (1998).
- Kürschner, W. M., van der Burgh, J., Visscher, H. & Dilcher, D. L. Oak leaves as biosensors of late Neogene and early Pleistocene paleoatmospheric CO₂ concentrations. *Mar. Micropaleontol.* **27**, 299–312 (1996).
- Veizer, J., Godderis, Y. & François, L. M. Evidence for decoupling of atmospheric CO₂ and global climate during the Phanerozoic eon. *Nature* **408**, 698–701 (2000).
- Thomas, E., Zachos, J. C. & Bralower, T. J. in *Warm Climates in Earth History* (eds Huber, B. T., MacLeod, K. G. & Wing, S. L.) 132–160 (Cambridge Univ. Press, Cambridge, 2000).
- Krull, E. S., Retallack, G. J., Campbell, H. J. & Lyon, G. L. $\delta^{13}C_{org}$ chemostratigraphy of the Permian–Triassic boundary in the Maitai Group, New Zealand: evidence for high-latitude methane release. *NZ J. Geol. Geophys.* **43**, 21–32 (2000).
- Meyen, S. V. *Fundamentals of Palaeobotany* (Chapman & Hall, London, 1987).
- Retallack, G. J. Postapocalyptic greenhouse revealed by earliest Triassic paleosols in the Sydney Basin, Australia. *Geol. Soc. Am. Bull.* **111**, 52–70 (1999).
- Anderson, J. M. & Anderson, H. M. *Palaeoflora of Southern Africa, Molteno Formation (Triassic)* Vol. 2 *Gymnosperms (excluding Dicotyledon)* (Balkema, Rotterdam, 1989).
- Gomankov, A. V. & Meyen, S. V. *Tatarinovaya flora (soslav i rasprostraneniye v pozdnei permi Evrazii) [Tatarian flora (composition and distribution in the late Permian of Eurasia)]*. (Trudy Akademii Nauk SSSR 401, 1986).
- Florin, R. Untersuchungen zur Stammesgeschichte der Coniferales und Cordaitales. *K. Svenska Vet. Akad. Handl.* **10**(1), 1–588 (1931).
- Gradstein, F. M. et al. in *Geochronology, Time Scales and Global Stratigraphic Correlation* (eds Berggren, W. A., Kent, D. V., Aubry, M.-P. & Hardenbol, J.) 95–126 (Spec. Pap. 54, Soc. Econ. Paleont. Mineral., Tulsa, 1995).
- Royer, D. L., Berner, R. A. & Hickey, L. J. Estimating latest Cretaceous and early Tertiary atmospheric pCO₂ from stomatal indices. *Geol. Soc. Am. Abstr.* **32**(7), A196 (2000).
- Lemarchand, D., Gaillardet, J., Lewin, E. & Allègre, C. J. The influence of rivers on marine boron isotopes and implications for reconstructing past ocean pH. *Nature* **408**, 951–954 (2000).
- Retallack, G. J., Bestland, E. A. & Fremd, T. Eocene and Oligocene paleosols in central Oregon. *Geol. Soc. Am. Spec. Pap.* **344**, 1–192 (2000).
- Vakrameev, V. A. *Jurassic and Cretaceous Floras and Climates of the Earth* 53 (Cambridge Univ. Press, Cambridge, 1991).
- Pálffy, J. et al. Timing the end-Triassic mass extinctions: first on land, then in the sea. *Geology* **28**, 39–42 (2000).
- McElwain, J. C., Beerling, D. J. & Woodward, F. I. Fossil plants and global warming at the Triassic–Jurassic boundary. *Science* **285**, 1386–1390 (1999).

22. Jahren, A. H., Arens, N. C., Sarmiento, G., Guerro, J. & Amundson, R. Terrestrial record of methane hydrate dissociation in the Early Cretaceous. *Geology* **29**, 159–162 (2001).
23. Arens, N. C. & Jahren, A. H. Carbon isotope excursion in atmospheric CO₂ at the Cretaceous-Tertiary boundary: evidence from terrestrial sediments. *Paleo* **15**, 314–322 (2000).
24. Khalil, M. A. K. (ed.) *Atmospheric Methane* 86 (Springer, Berlin, 2000).
25. McGowan, B. & Li, Q.-Y. Miocene climatic oscillation recorded in the Lakes Entrance oil shaft, southern Australia. *Aust. J. Earth Sci.* **43**, 129–148 (1997).
26. Utescher, T., Mossbruger, U. & Ashraf, A. R. Terrestrial climate evolution in northwest Germany over the last 25 million years. *Paleo* **15**, 430–449 (2000).
27. Schwartz, T. Lateritic bauxite in central Germany and implications for Miocene palaeoclimates. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **129**, 37–50 (1997).
28. Cowling, S. A. Plants and temperature: CO₂ uncoupling. *Science* **285**, 1500–1501 (1999).
29. McElwain, J. C. & Chaloner, W. C. The fossil cuticle as a skeletal record of environmental change. *Paleo* **11**, 376–388 (1996).
30. Retallack, G. J. *A Colour Guide to Paleosols* 117 (Wiley, Chichester, 1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

A. Liston provided herbarium specimens of *Ginkgo*, M. Shaffer helped with SEM imaging, and S. Tanaka supplied Japanese literature. W. Kürschner, D. Dilcher, S. Schekler and V. Wilde offered useful botanical discussion. M. Manga and J. Wynn helped with curve fitting and diffusion equations.

Correspondence and requests for materials should be addressed to the author (e-mail: greg@darkwing.uoregon.edu).

Strong coherence between solar variability and the monsoon in Oman between 9 and 6 kyr ago

U. Neff*, S. J. Burns†‡, A. Mangini*, M. Mudelsee§, D. Fleitmann† & A. Matter†

* Heidelberg Academy of Sciences, Im Neuenheimer Feld 229, Heidelberg, Germany D-69120

† Geological Institute, University of Bern, Baltzerstrasse 1, Bern, Switzerland CH-3012

§ Institute of Meteorology, University of Leipzig, Stephanstrasse 3, Leipzig, Germany D-04103

Variations in the amount of solar radiation reaching the Earth are thought to influence climate, but the extent of this influence on timescales of millennia to decades is unclear. A number of climate records show correlations between solar cycles and climate¹, but the absolute changes in solar intensity over the range of decades to millennia are small² and the influence of solar flux on climate is not well established. The formation of stalagmites in northern Oman has recorded past northward shifts of the intertropical convergence zone³, whose northward migration stops near the southern shoreline of Arabia in the present climate⁴. Here we present a high-resolution record of oxygen isotope variations, for the period from 9.6 to 6.1 kyr before present, in a Th–U-dated stalagmite from Oman. The $\delta^{18}\text{O}$ record from the stalagmite, which serves as a proxy for variations in the tropical circulation and monsoon rainfall, allows us to make a direct comparison of the $\delta^{18}\text{O}$ record with the $\Delta^{14}\text{C}$ record from tree rings⁵, which largely reflects changes in solar activity^{6,7}. The excellent correlation between the two records suggests that one of the primary controls on centennial- to decadal-scale changes in tropical rain-

fall and monsoon intensity during this time are variations in solar radiation.

The Indian Ocean monsoon is one of the main weather systems on Earth, and variations in its intensity have broad oceanographic and economic effects. To date, studies of how and why the monsoon varies through time have been restricted to studies of meteorological records, which extend back about 150 years⁸, or studies of lacustrine and marine sediments^{9–12}, which mainly yield information on millennial and longer timescales. At present, information on monsoon variation on decadal to centennial timescales is limited to identification of 'centennial-scale' changes in sedimentary records^{13,14}, but these studies lack the resolution to determine specific periodicities or forcing mechanisms. We have developed a high-resolution proxy for estimating variation in monsoon intensity by measuring past changes in $\delta^{18}\text{O}$ of monsoon rainfall as recorded in calcite $\delta^{18}\text{O}$ of a stalagmite.

Hoti cave is located in northern Oman on the southwestern side of the Oman mountains (57° 21' E, 23° 05' N, 800 m above sea level). The modern climate of the area is arid to semi-arid, and the area is not at present affected by the Indian Ocean monsoon system. However, numerous marine and continental palaeoclimate records indicate that the summer monsoon was considerably stronger during early to middle Holocene times than it is at present^{9–12,15}, accompanied by a shift in the northern limit of the monsoon rainfall belt far north of its modern location^{15,16}, which resulted in a continental pluvial period in the Sahel region of Africa, in Arabia and in India. In Hoti cave³, the pluvial period led to deposition of a set of large stalagmites.

Here we present a high-resolution profile of $\delta^{18}\text{O}$ of stalagmite H5 from Hoti cave³ on an improved timescale dated with 12 Th–U ages measured using mass spectrometry (see Methods for details). The speleothem covers the time span from 9.6 to 6.1 kyr before present (BP). It can be divided into four sections exhibiting different growth rates ranging from 0.03 to 0.57 mm yr^{−1} (Fig. 1). Samples (826) for stable isotope measurements were drilled from the core by hand with an average sampling interval of 0.4 mm, yielding an average time resolution for the core of 4.1 years. For the fastest growth interval (high-resolution interval), however, between 7.9 and 8.3 mm yr^{−1}, the resolution increases to 1.4 years.

The $\delta^{18}\text{O}$ values vary between −4‰ and −6‰ VPDB, reflecting the characteristically low $\delta^{18}\text{O}$ of monsoonal rainfall¹⁷ (Fig. 2a). The values of the upper 4 mm, which are not displayed in Fig. 2a, increase to −1.5‰, marking the end of the Holocene pluvial period. Modern stalagmites in this cave also have comparatively positive $\delta^{18}\text{O}$ ranging from 0‰ to −2‰, corresponding to the present arid climate³. Measurements of $\delta^{18}\text{O}$ of modern cave waters and actively

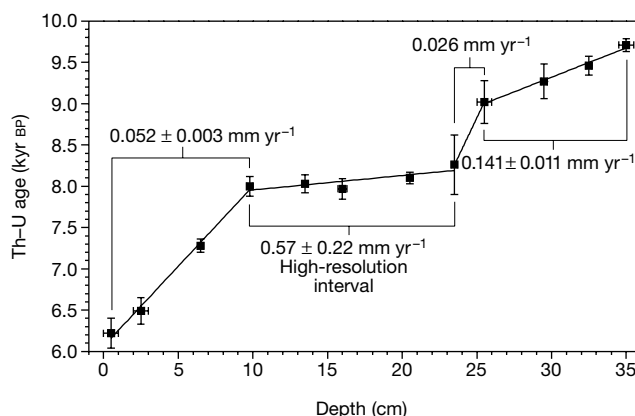


Figure 1 Plot of age versus depth for stalagmite H5. Twelve Th–U ages identify four sections with different growth rates varying between 0.03 and 0.57 mm yr^{−1}. All errors are 2σ (see Supplementary Information).

‡ Present address: Department of Geosciences, University of Massachusetts, Amherst, Massachusetts 01003, USA.

growing stalagmites show that calcite precipitation takes place in isotopic equilibrium³, which was probably also the case for the wetter climate of early to middle Holocene times, when the relative humidity in the cave would have been greater. The variation of the $\delta^{18}\text{O}$ signal is very unlikely to be directly related to temperature changes because mean annual temperatures in the tropics have varied less than 1 °C during the Holocene¹⁸, which could account for only 0.25‰ variation in calcite $\delta^{18}\text{O}$. Nor are the $\delta^{18}\text{O}$ values of precipitation in tropical areas correlated with temperature¹⁷. Kinetic effects on the $\delta^{18}\text{O}$ values¹⁹ also seem unlikely because there is no observed relationship between H5 growth rates, which vary by a factor of 10, and its isotopic ratios. The $\delta^{18}\text{O}$ values of monsoonal rainfall and other types of rainfall associated with the ITCZ do, however, show an inverse correlation with rainfall amount^{17,20}. We, therefore, ascribe the variation in $\delta^{18}\text{O}$ to changing rainfall intensity. Peaks of heavier $\delta^{18}\text{O}$ are associated with diminished monsoonal precipitation, due to either periodic southwards shifts of the ITCZ or periods of weaker monsoon, though it should be noted that the location of the ITCZ is not rigidly coupled to the variations of monsoon rainfall. As the $\delta^{18}\text{O}$ values of rainfall and ground water vary with changing intensity of monsoon rainfall, these variations are recorded in speleothem calcite.

The high temporal resolution of the speleothem $\delta^{18}\text{O}$ record enables a precise comparison with atmospheric $\Delta^{14}\text{C}$ measured in tree-rings⁵ (Fig. 2a and b). This was done by further refining the Th–U timescale by tuning it to the $\Delta^{14}\text{C}$ timescale. We emphasize

that the tuning was done while always remaining within the measurement error of each of the 12 Th–U ages along the profile, so that the net shift of any part of the $\delta^{18}\text{O}$ curve is only 190 years at most (Fig. 3). The similarity between the smoothed $\delta^{18}\text{O}$ and $\Delta^{14}\text{C}$ time series, both in their general patterns and in the number of peaks, is extremely strong. Even millennial-scale trends and relative amplitudes correspond. Furthermore, the high-resolution interval between 7.9 and 8.3 kyr BP also reveals a close correspondence between the two curves. The parallel evolution of $\delta^{18}\text{O}$ and $\Delta^{14}\text{C}$ seems very unlikely to have occurred by chance. Rather, the high correlation provides solid evidence that both signals are responding to the same forcing. Variations of $\Delta^{14}\text{C}$ were attributed to changes in the production rate in the stratosphere, induced by solar wind modulation of the cosmic ray flux. Maxima of ^{10}Be concentrations in polar ice cores that are synchronous with maxima in $\Delta^{14}\text{C}$ further reinforce this interpretation^{6,7,21}.

The high resolution and dating precision of the $\delta^{18}\text{O}$ record of H5 make it possible to perform a reliable frequency analysis. Spectral analyses of the untuned $\delta^{18}\text{O}$ record are given in Fig. 4a and b. The $\delta^{18}\text{O}$ results show statistically significant periodicities centred on 1,018, 226, 28, 10.7 and 9 years. Two broader sets of cycles are centred between 101–90 years and 35–26 years. These cycles are close to the periodicities of the tree-ring $\Delta^{14}\text{C}$ record (206, 148, 126, 89, 26 and 10.4 years), which are assigned to solar modulation⁷. Spectral analyses of $\delta^{18}\text{O}$ values of foraminifera from sediments in the South China Sea also suggest monsoon cycles of around 1,000 (775) and 100–90 years (ref. 22).

We performed an additional spectral analysis of the tuned $\delta^{18}\text{O}$ record (Fig. 4c and d). The shape of the tuned spectra is rather similar to that of the untuned spectra. The tuned time series is somewhat shorter than the untuned one. Therefore all peaks of the tuned spectra are moved to somewhat shorter periods. Owing to the tuning, the noise between the significant peaks decreases. Statistically significant periodicities are centred on 779, 205, 134 and 87 years. The latter cycle is particularly strengthened in the tuned spectral analysis, as the two separate cycles at 101 and 90 years are joined in one stronger peak. The broad set of peaks around 28–23 years is present in both analyses as well as the two peaks at 10 and 9 years. Spectral analysis of the tuned high-resolution interval reveals further cycles with periodicities of 24, 7.5 and 3 years. Perhaps not surprisingly, the tuned timescale results in a cycle pattern that even more closely matches the periodicities of $\Delta^{14}\text{C}$ (206, 148, 126, 89, 26 and 10.4 years)⁷. The cross-spectral analysis (Fig. 5) of these two time series further confirms the correspondence

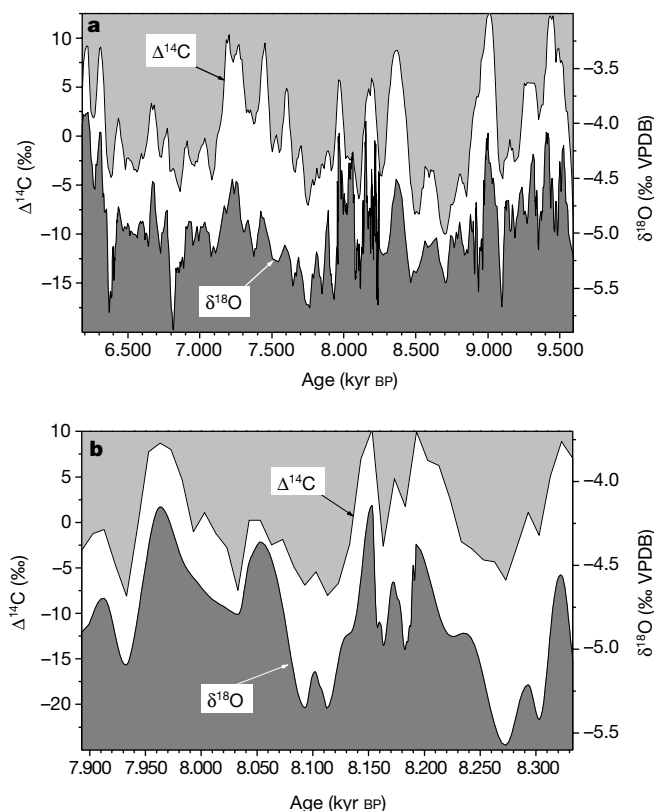


Figure 2 Profiles of H5 $\delta^{18}\text{O}$ values and atmospheric $\Delta^{14}\text{C}$. **a**, The entire H5 record (826 samples); **b**, the high-resolution interval. Both profiles in **a** were smoothed with 5-point adjacent averaging for better visual comparison. The $\delta^{18}\text{O}$ profile in **b** was filtered with 7-point fast Fourier transform smoothing (cut-off frequency, 0.1 yr^{-1}). The correlation coefficient of the unsmoothed data is $r = 0.60$, $P(>|r|) < 10^{-8}$ in **a**, and $r = 0.55$, $P(>|r|) = 1.1 \times 10^{-4}$ in **b**. Because of the apparent good relationship between the two profiles, we fine-tuned the peaks of the $\delta^{18}\text{O}$ age profile to the peaks of the $\Delta^{14}\text{C}$ record. The corrections of the Th–U timescale are shown in Fig. 3.

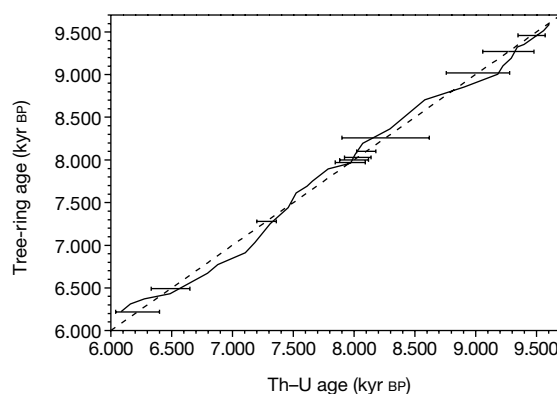


Figure 3 Measured (dashed line) and optimized (solid line) Th–U age scales for stalagmite H5. The error bars are the uncertainties of the Th–U ages. The correlation between the time series of $\delta^{18}\text{O}$ and $\Delta^{14}\text{C}$ was made using the adjusted Th–U timescale. As shown, the optimized timescale always remains within the uncertainty of the individual Th–U dates, both for the entire profile of $\delta^{18}\text{O}$ and for the high-resolution interval.

of the periodicities. This similarity suggests a solar forcing of the Hoti cave palaeoprecipitation record, and by extension, of the Indian Ocean monsoon, during this period.

The periodicities in the high-resolution interval are similar to the cycles of 24, 7.5 and 3.2 years that have been noted for the North Atlantic Oscillation²³, suggesting that at these shorter timescales the North Atlantic climate system might also be a cause of changes in the amount of monsoon rainfall in the Hoti cave area. In particular, it has been suggested that the 512-year cycle in $\Delta^{14}\text{C}$, which is within the bandwidth around the 779-year cycle in the H5 record, is caused by variations in ocean circulation⁷—specifically, by changes in North Atlantic Deep Water (NADW) formation that control incorporation of ^{14}C into the ocean²⁴. Periods of decreased NADW formation (decreased surface heat flux to the North Atlantic Ocean) should be associated with longer, colder Eurasian winters, leading to increased snow cover on the Tibetan plateau and a delayed or and weakened monsoon the following summer^{25,26}.

In summary, the $\delta^{18}\text{O}$ record of the stalagmite H5 from Hoti cave

represents a precisely dated high-resolution time series of the intensity of Indian Ocean monsoonal rainfall. The strong similarity between the smoothed secular variation curves of the H5 $\delta^{18}\text{O}$ isotopic record and the $\Delta^{14}\text{C}$ record—together with their spectral analyses—suggest that both are responding to the same climate forcing. Much of the variation in $\Delta^{14}\text{C}$ is attributed to solar forcing through variations in solar activity and intensity⁷. Thus, to the extent that centennial and decadal variations in the $\Delta^{14}\text{C}$ record are controlled by changes in global solar radiation, so should such changes control the intensity of the Indian Ocean monsoon. Less clear is the mechanism by which solar variability affects the monsoon. The variations in solar irradiance necessary to cause the observed changes in $\Delta^{14}\text{C}$ are probably of the order of a few tenths of one per cent (ref. 2). Such minor variations are unlikely to have directly caused significant differences in sensible heating of the Tibetan plateau. It is more likely that solar variability leads to changes in atmospheric or oceanic circulation that amplify this initial input. For example, a model study of the relationship between the 11-year solar cycle and climate has suggested that maximum solar irradiance leads to increased easterly trade winds and a broadened Hadley cell²⁷. □

Methods

Th–U ages

Analytical procedures for the separation and purification of Th and U were as described in ref. 28. Th–U measurements were performed on a multicollector mass spectrometer (Finnigan MAT 262 RPQ) with a double filament technique. Thorium and uranium were measured in peak-jump mode and semi-peak-jump mode, respectively. Calibration of Faraday cup to ionization efficiency was made adopting the natural $^{238}\text{U}/^{235}\text{U}$ of 137.88. To determine the uranium and thorium concentrations, defined quantities of a $^{233}\text{U}/^{236}\text{U}$ double spike and a ^{229}Th spike were added. Th–U ages were corrected for detritus following ref. 29 assuming a $^{232}\text{Th}/^{238}\text{U}$ isotope ratio of 3.8. The reproducibility of the isotope ratio of $^{234}\text{U}/^{238}\text{U}$ and the concentration of ^{232}Th of standard materials is 0.3% and 0.8% (2σ), respectively. For details about measurements of standard material, see ref. 29.

Oxygen isotopes

Oxygen isotope ratios are expressed in δ -notation, the per mil deviation from the VPDB standard: $\delta^{18}\text{O} = [((^{18}\text{O}/^{16}\text{O})_{\text{sample}} / (^{18}\text{O}/^{16}\text{O})_{\text{VPDB}}) - 1] \times 1,000$. For each measurement, approximately 3 mg of powder was drilled from the sample and analysed with an on-line, automated, carbonate preparation system linked to a VG Prism ratio mass spectrometer. Reproducibility of standard materials is 0.08‰ for $\delta^{18}\text{O}$.

Received 14 September 2000; accepted 26 March 2001.

- Hoyt, D. V. & Schatten, K. H. *The Role of the Sun in Climate Change* (Oxford Univ. Press, New York, 1997).
- Fligge, M. & Solanki, S. K. The solar spectral irradiance since 1700. *Geophys. Res. Lett.* **26**, 2465–2468 (1999).
- Burns, S. J., Matter, A., Frank, N. & Mangini, A. Speleothem-based paleoclimate record from northern Oman. *Geology* **26**, 499–502 (1998).
- Hastenrath, S. *Climate and Circulation of the Tropics* (Reidel, Boston, 1985).
- Stuiver, M. et al. INTCAL98 Radiocarbon age calibration, 24,000–0 cal BP. *Radiocarbon* **40**, 1041–1083 (1998).
- Beer, J., Mende, W. & Stellmacher, R. The role of the sun in climate forcing. *Quat. Sci. Rev.* **19**, 403–415 (2000).
- Stuiver, M. & Braziunas, T. F. Sun, ocean climate and atmospheric $^{14}\text{CO}_2$: an evaluation of causal and spectral relationships. *Holocene* **3**, 289–305 (1993).
- Sontakke, N. A., Plant, G. B. & Singh, N. Construction of all-India summer monsoon rainfall series for the period 1844–1991. *J. Clim.* **6**, 1807–1811 (1993).
- Clemens, S. C., Murray, D. W. & Prell, W. L. Nonstationary phase of the Plio-Pleistocene Asian monsoon. *Science* **274**, 943–948 (1996).
- Overpeck, J., Anderson, D., Trumbore, S. & Prell, W. The southwest Indian Monsoon over the last 18 000 years. *Clim. Dyn.* **12**, 213–225 (1996).
- Bryson, R. A. & Swain, A. M. Holocene variations of monsoon rainfall in Rajasthan. *Quat. Res.* **16**, 135–145 (1981).
- Gasse, F., Teher, R., Durand, A., Gibert, E. & Fontes, J.-C. The arid-humid transition in the Sahara and the Sahel during the last deglaciation. *Nature* **346**, 141–146 (1990).
- Lamb, H. F. et al. Relation between century-scale Holocene arid intervals in tropical and temperate zones. *Nature* **373**, 134–137 (1995).
- Sirocko, F. et al. Century-scale events in monsoonal climate over the last 24,000 years. *Nature* **364**, 322–324 (1993).
- Liu, K., Yao, Z. & Thompson, L. G. A pollen record of Holocene climatic changes from the Dunde ice cap, Qinghai–Tibetan Plateau. *Geology* **26**, 135–138 (1998).
- Ritchie, J. C., Eyles, C. H. & Haynes, C. V. Sediment and pollen evidence for an early to mid-Holocene humid period in the eastern Sahara. *Nature* **314**, 352–355 (1985).

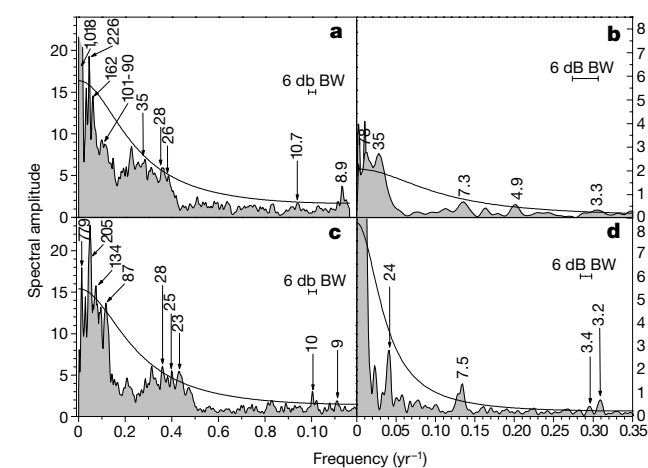


Figure 4 Frequency analysis of the H5 $\delta^{18}\text{O}$ record for the entire sample and for the high-resolution interval. **a, b**, Untuned; **c, d**, tuned. Peaks are labelled with their period (in years). Spectral power was estimated with the SPECTRUM package³⁰, which uses the Lomb–Scargle periodogram for unevenly spaced data. The number (n_{seg}) of overlapping (50%) segments is 6 (**a** and **c**) and 3 (**b** and **d**). The spectral window is a Welch I type. The 6-dB bandwidth (BW) determines the frequency resolution. Smooth line, red-noise background (upper 90% χ^2 bound).

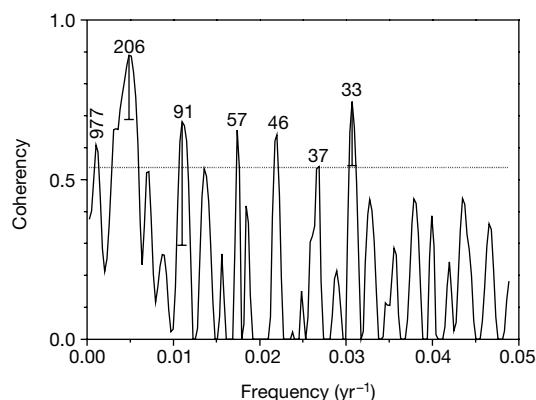


Figure 5 Cross-spectral analysis³⁰ of the tuned H5 $\delta^{18}\text{O}$ record for the entire studied sample versus the $\Delta^{14}\text{C}$ record (ref. 4). The same spectral parameters ($n_{\text{seg}} = 6$, Welch I window) as for the univariate (Fig. 4c) analysis were used. Peaks common to both time series are labelled with their period (in years). Error bars are given as lower 95% bounds. Horizontal line is 5% false-alarm level³⁰.

17. Rozanski, K., Araguás-Araguás, L. & Gonfiantini R. in *Climate Change in Continental Isotopic Records* (eds Swart, P. K., Lohmann, K. C., McKenzie, J. & Savin, S.) 1–36 (American Geophysical Union Geophysical Monograph 78, Washington DC, 1993).
18. COHMAP Members. Climatic changes of the last 18 000 years: observations and model simulations. *Science* **241**, 1043–1052 (1988).
19. Hendy, C. H. The isotopic geochemistry of speleothems - I. The calculation of the effects of different modes of formation on the isotopic composition of speleothems and their applicability as paleoclimate indicators. *Geochim. Cosmochim. Acta* **35**, 801–824 (1971).
20. Technical Report Series No. 331, 781 (International Atomic Energy Agency, Vienna, 1992).
21. Lean, J., Beer, J. & Bradley, R. Reconstruction of solar irradiance since 1610: Implications for climate change. *Geophys. Res. Lett.* **22**, 3195–3198 (1995).
22. Wang, L. *et al.* East Asian monsoon climate during the Late Pleistocene: high resolution sediment records from the South China Sea. *Mar. Geol.* **156**, 245–284 (1999).
23. Cook, E. R., D'Arrigo, R. D. & Briffa, K. R. A reconstruction of the North Atlantic oscillation using tree-ring chronologies from North America and Europe. *Holocene* **8**, 9–17 (1998).
24. Stocker, T. F. & Wright, D. G. Rapid changes in ocean circulation and atmospheric radiocarbon. *Paleoceanography* **11**, 773–795 (1996).
25. Dickson, R. R. Eurasian snow cover versus Indian monsoon rainfall—An extension of the Hahn-Shukla results. *J. Clim. Appl. Meteorol.* **23**, 171–173 (1984).
26. Meehl, G. A. Influence of the land surface in the Asian summer monsoon, external conditions versus internal feedbacks. *J. Clim.* **7**, 1033–1049 (1994).
27. Haigh, J. D. The impact of solar variability on climate. *Science* **272**, 981–984 (1996).
28. Ivanovich, M. & Harmon, R. S. *Uranium Series Disequilibrium: Applications to Environmental Problems* (Clarendon, Oxford, 1993).
29. Frank, N., Braun, M., Hambach, U., Mangini, A. & Wagner, G. Warm period growth of travertine during the last interglaciation in southern Germany. *Quat. Res.* **54**, 38–48 (2000).
30. Schulz, M. & Stettin, K. Spectral analysis of unevenly spaced paleoclimatic time series. *Comput. Geosci.* **23**, 929–945 (1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank D. Sanz for caving assistance; R. Eichstädter for technical assistance; M. Stuiver, A. Baker and D. Ford for suggestions; and S. Clemens for comments.

Correspondence and requests for materials should be addressed to S.B. (e-mail: sburns@geo.umass.edu).

Metamorphic devolatilization of subducted marine sediments and the transport of volatiles into the Earth's mantle

D. M. Kerrick* & J. A. D. Connolly†

* Department of Geosciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

† Earth Sciences Department, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland

Volatiles, most notably CO₂, are recycled back into the Earth's interior at subduction zones^{1,2}. The amount of CO₂ emitted from arc volcanism appears to be less than that subducted, which implies that a significant amount of CO₂ either is released before reaching the depth at which arc magmas are generated or is subducted to deeper depths. Few high-pressure experimental studies^{3–5} have addressed this problem and therefore metamorphic decarbonation in subduction zones remains largely unquantified, despite its importance to arc magmatism, palaeo-atmospheric CO₂ concentrations and the global carbon cycle⁶. Here we present computed phase equilibria to quantify the evolution of CO₂ and H₂O through the subduction-zone metamorphism of carbonate-bearing marine sediments (which are considered to be a major source for CO₂ released by arc volcanoes⁶). Our analysis indicates that siliceous limestones undergo negligible

devolatilization under subduction-zone conditions. Along high-temperature geotherms clay-rich marls completely devolatilize before reaching the depths at which arc magmatism is generated, but along low-temperature geotherms, they undergo virtually no devolatilization. And from 80 to 180 km depth, little devolatilization occurs for all carbonate-bearing marine sediments. Infiltration of H₂O-rich fluids therefore seems essential to promote subarc decarbonation of most marine sediments. In the absence of such infiltration, volatiles retained within marine sediments may explain the apparent discrepancy between subducted and volcanic volatile fluxes and represent a mechanism for return of carbon to the Earth's mantle.

A premise of our work is that realistic modelling of metamorphic devolatilization of subducted lithologies is only possible on the basis of phase equilibria in chemical systems closely approximating actual bulk compositions. Our studies on metamorphic devolatilization of the other two main carbonate-bearing lithologies involved in subduction zones (ophicarbonates and metabasalts) are considered elsewhere^{2,7}. Carbonate is abundant in two main pelagic marine sediment lithologies⁸: (1) siliceous limestones and (2) clay-carbonates (marls). From the database of ref. 8 (see Supplementary Information) we selected bulk compositions of siliceous limestones from the Marianas and Vanuatu trenches, a marl from the Antilles trench and their average marine sediment bulk composition (denoted 'GLOSS' in ref. 8). Our computations account for the oxide components: SiO₂, Al₂O₃, FeO, MgO, CaO, Na₂O, K₂O, CO₂ and H₂O.

For each marine sediment bulk composition, the corresponding phase diagram section (Fig. 1) was computed as a function of pressure (*P*) and temperature (*T*) by free-energy minimization⁹. The thermodynamic database of ref. 10 was used for the properties of all end-member species, and mineral solutions were modelled as described elsewhere⁹. Thermodynamic data for H₂O, CO₂ and their mixtures were computed from the equation of state given in ref. 11. To track metamorphic devolatilization along the top of subducted slabs, we adopted the geotherms¹² for the subduction zones of northwestern and southeastern Japan (Fig. 1). These geotherms are reasonable approximations for the respective extremal low-temperature and high-temperature geotherms for subduction zones (S. M. Peacock, personal communication).

Because of compositional degrees of freedom in the crystalline and fluid phases, the phase diagram sections are dominated by multivariant phase fields. Consequently, both mineral modes and compositions vary continuously along geotherms (Fig. 2). Phase relations along geotherms up to a pressure *P* ≈ 3 GPa change significantly (Fig. 1) because of intersection with numerous phase field boundaries. However, at *P* > 3 GPa, the geotherms are subparallel to the phase field boundaries (Fig. 1); consequently, little reaction occurs along geotherms at *P* > 3 GPa. The differences between these regimes are illustrated in Fig. 2. Accordingly, significant changes in the mineralogy and mineral proportions occur up to ~800 °C (*P* ≈ 3 GPa) whereas there is comparatively little variation above ~800 °C.

The fluid composition (Fig. 2) is controlled by multivariant equilibria involving carbonates and hydrous phases. The rise in the mole fraction of CO₂, *X*_{CO₂}, up to ~750 °C correlates with consumption of carbonates (Fig. 2), whereas the diminution in *X*_{CO₂} above ~750 °C occurs because of aragonite production. To track loss of volatiles we computed the percentage (by weight; wt%) of H₂O and CO₂ for carbonate-bearing marine sediments as a function of pressure and temperature (Fig. 3). In the lower-pressure half of Fig. 3, the negative *P*–*T* slopes of the wt% H₂O isopleths reflect negative slopes of phase field boundaries (Fig. 1). In contrast to isopleths with negative slopes at lower pressures, isopleths are subparallel to geotherms at *P* > 2–3 GPa (Fig. 3).

Siliceous limestones release about 1 wt% CO₂ and 1 wt% H₂O along the high-temperature geotherm (Fig. 3c, d). Because less CO₂

and H₂O are released along lower-temperature geotherms, most of the volatile content of siliceous limestones is retained to depths of 180 km and thus such lithologies would undergo little devolatilization upon subduction. This conclusion is compatible with the existence of ultrahigh-pressure marbles¹³.

In contrast with siliceous limestones, H₂O-rich lithologies with low initial carbonate contents (that is, clay-rich marls) are predicted to undergo considerably more devolatilization (Fig. 3b). Along the high-temperature geotherm, all of the initial CO₂ (4 wt%) and most of the initial H₂O (10–11 wt%) is released by 90 km depth (that is, forearcs). For geotherms in the lower-temperature half of the area bounded by the limiting geotherms (Fig. 3b), relatively little CO₂ and H₂O would be released. For various geotherms in the higher-temperature half of the area bounded by the limiting geotherms (Fig. 3b), there are significant differences in the amount of devolatilization.

For the carbonate-bearing protoliths considered here, the geotherms at 80–180 km are subparallel to the H₂O and CO₂ isopleths (Fig. 3); thus, little or no devolatilization is expected. Consequently, for closed-system behaviour, subducted carbonate-bearing marine sediments would not provide a source of volatiles for arc magmatism. However, decarbonation of marine sediments at these depths may be driven by infiltration of H₂O-rich fluids originating from intercalated hydrous pelagic or terrigenous sediments, and/or metabasalts in the subjacent slab. Computed⁷ and experimentally determined¹⁴ high-pressure phase equilibria imply that significant proportions of the initial H₂O in subducted oceanic metabasalts are released under forearcs and subarcs. Infiltration of the evolved fluid into the overlying subducted sediments would induce decarbonation. But because there are no major dehydration ‘pulses’ in subducted metabasalts under volcanic arcs^{7,14}, no corresponding pervasive infiltration of water from dehydrating meta-

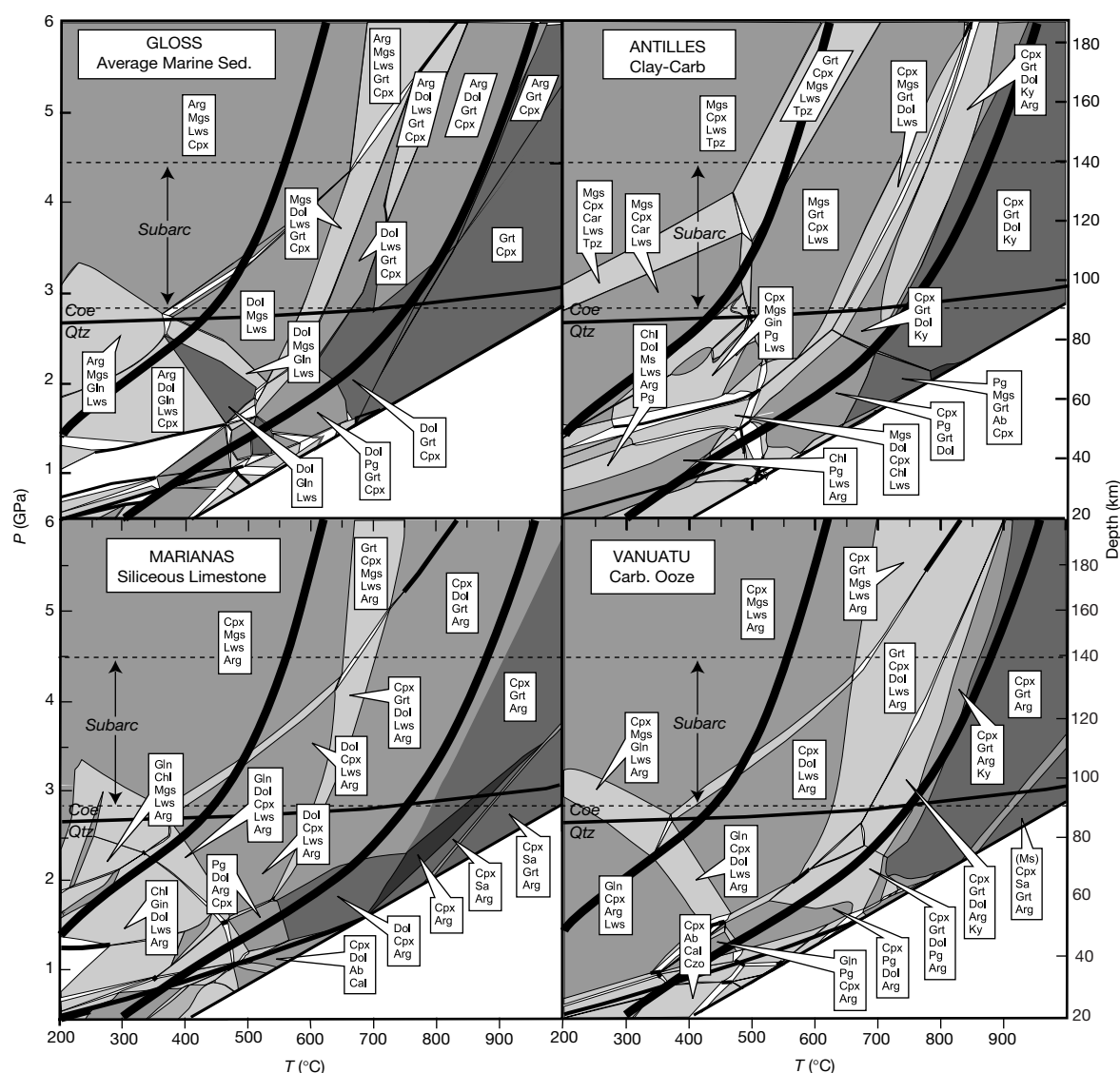


Figure 1 Phase equilibria computed for selected bulk compositions from the database of ref. 8. Further details are available in Supplementary Information. The computer programs, and the thermodynamic data and equations of state utilized by these programs, are available at www.erdw.ethz.ch/~jamie/perplex.html. Mineral abbreviations are: Ab, albite; Arg, aragonite; Cal, calcite; Car, Mg-Fe carpholite; Chl, chlorite; Coe, coesite; Cpx, clinopyroxene; Czo, clinozoisite; Dol, dolomite; Gln, glaucophane; Grt, Garnet; Ky, kyanite; Lws, lawsonite; Mgs, magnesite; Pg, paragonite; Qtz, quartz; Sa, Sanidine; Tpz, topaz.

With the exception of muscovite absence (denoted by Ms) in a high-temperature field of Vanuatu, muscovite + quartz/coesite + fluid is present in all phase fields. The shading denotes the variance of the phase fields. Univariant phase fields are denoted by thick lines. The semi-parallel thickest curves are geotherms for southeastern (right) and northwestern (left) Japan¹². The subarc depth range is from ref. 18. Assemblages in the small phase fields below ~1.5 GPa are omitted for clarity.

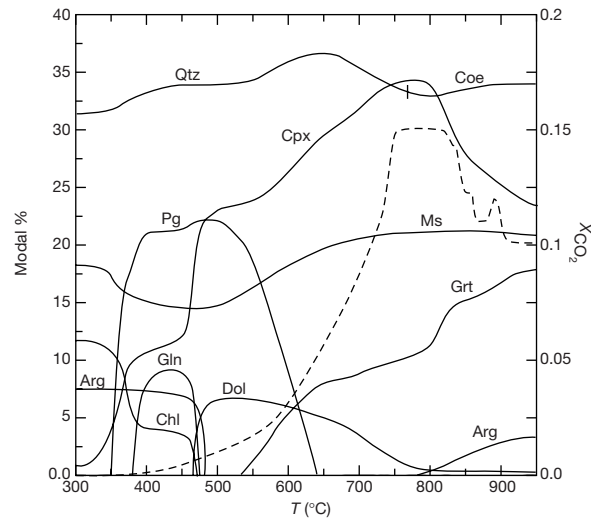


Figure 2 Modal percentages of minerals and fluid composition. Modal percentages of minerals (left ordinate and solid lines) and fluid composition (right ordinate and dashed line) are shown for the average marine sediment bulk composition ('GLOSS' in ref. 8)

along the high-temperature geotherm shown in Fig. 1. Phase abbreviations as in Fig. 1. The vertical line at $\sim 770^\circ\text{C}$ marks the quartz-coesite equilibrium.

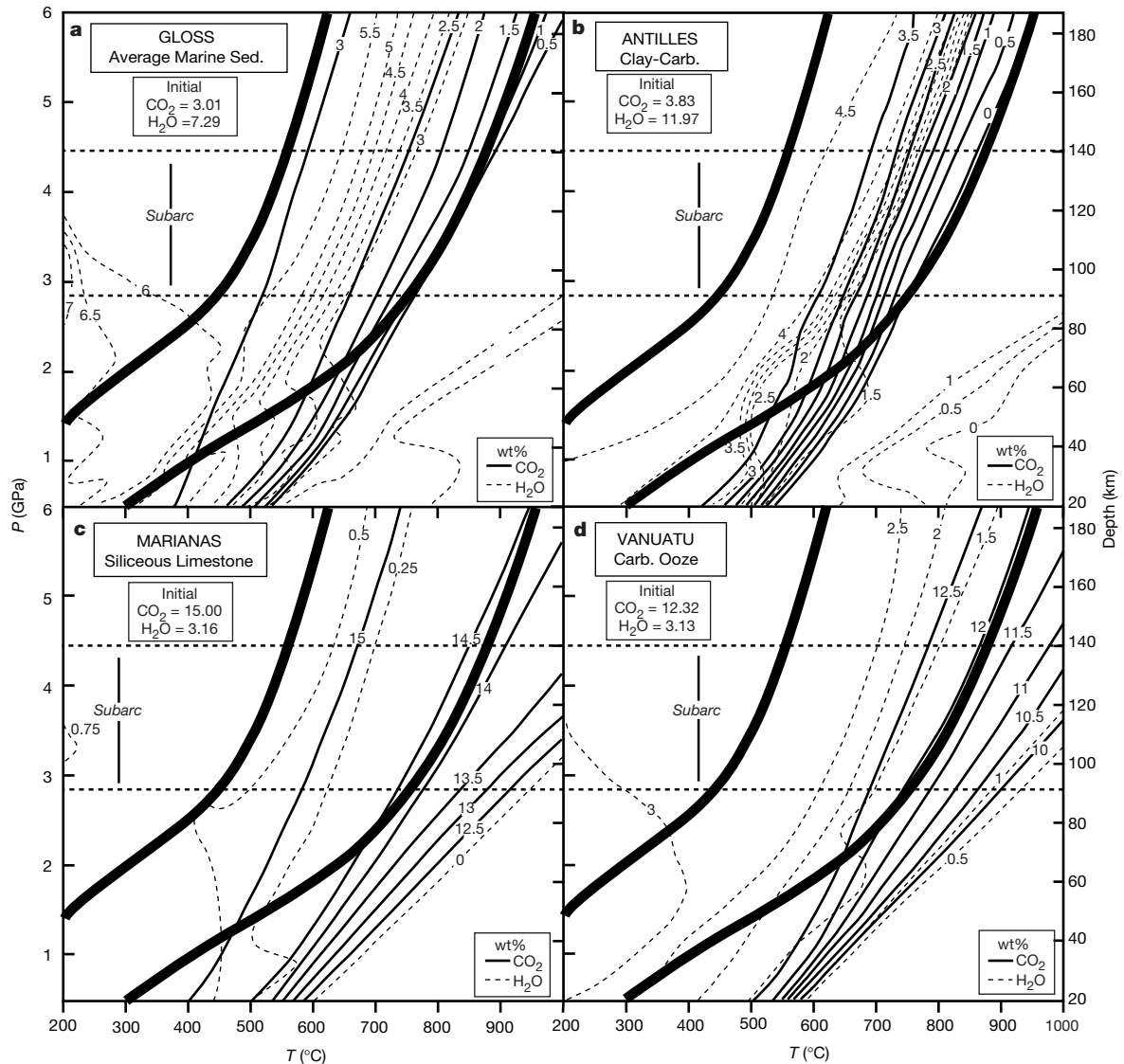


Figure 3 Weight percentages of CO_2 and H_2O for selected marine sediment bulk rock compositions (see Fig. 1). **a**, Gloss; **b**, Antilles; **c**, Marianas; **d**, Vanuatu. Heavy curved lines are limiting geotherms (see Fig. 1). Values of the initial wt% CO_2 and H_2O are given in the

insets (from ref. 8). The CO_2 and H_2O contents of the fluid phase can be determined by subtracting the data in these diagrams from the initial volatile contents of the protoliths.

Table 1 Subduction zone carbon budget

Subducted carbon (Tmol yr ⁻¹)*	
Sediment carbonate ⁸	1.2
Sediment organic carbon ¹	0.8
Oceanic metabasalts ¹⁹	3.4
Total	5.4
Expelled carbon (Tmol yr ⁻¹)	
Arc magmatism ²⁰	2–3
Carbon imbalance† (Tmol yr ⁻¹)	2.5–3.5

*1 Tmol = 10¹² mol.

† (Subducted carbon) – (expelled carbon).

basalts is expected in subarcs. Barring extensive infiltration of externally derived fluids, our study implies marked devolatilization under forearcs (for clay-rich marls with high-temperature geotherms) or retention of H₂O and CO₂ to depths well beyond subarcs (for siliceous limestones in all geotherms and clay-rich marls with low-temperature geotherms). Accordingly, most of the initial CO₂ and H₂O in subducted marine sediments will not be released beneath volcanic arcs. This inference is consistent with both the deficiency in the amount of CO₂ released from arc volcanoes compared to the amount of CO₂ contained within subducted carbonates (Table 1) and with the imbalance between subducted versus expelled H₂O (ref. 1).

Our equilibrium analysis implicitly assumes that there is no significant kinetic overstepping and metastability of metamorphic reactions. Although significant disequilibrium has been suggested for the transformation of anhydrous oceanic basalts and gabbros to eclogites¹⁵, the catalytic effect of H₂O (ref. 15) implies that equilibrium is more likely in dehydrating systems such as subducted sediments.

Melting is an alternative mechanism for release of volatiles from subducted sediment. Recent experiments using marine red clay¹⁶ suggest that sediment melting does not occur for the geotherms that we consider here. However, because metastable starting materials (for example, red clay) are unsuitable models for subduction-zone metamorphism and melting, confirmation of this conclusion requires experiments with more realistic initial mineral assemblages. Dissolution of minerals in supercritical fluids remains a possible, albeit largely unquantified, alternative mechanism for devolatilization.

As shown in Fig. 2, fluids produced by metamorphism of subducted marine sediments are H₂O-rich. Consequently, expulsion of such fluids to the overlying mantle wedge would not substantially affect the *P*–*T* conditions of melting (solidus) of the mantle wedge compared to those expected in the presence of a pure H₂O fluid.

Devolatilization of subducted sediment could contribute to seismicity along the tops of subducted slabs. The continuous nature of devolatilization is compatible with the spread of earthquake hypocentres along individual subduction zones¹⁷. However, correlation of slab seismicity with metamorphic devolatilization of subducted sediments needs to consider the marked differences in devolatilization for different bulk compositions and geotherms. □

Received 7 June 2000; accepted 28 February 2001.

1. Bebout, G. E. The impact of subduction-zone metamorphism on mantle-ocean chemical cycling. *Chem. Geol.* **126**, 191–218 (1995).
2. Kerrick, D. M. & Connolly, J. A. D. Subduction of ophiocarbonates and recycling of CO₂ and H₂O. *Geology* **26**, 375–378 (1998).
3. Yaxley, G. M. & Green, D. H. Experimental demonstration of refractory carbonate-bearing eclogite and siliceous melt in the subduction regime. *Earth Planet. Sci. Lett.* **128**, 313–325 (1994).
4. Molina, J. F. & Poli, S. Carbonate stability and fluid composition in subducted oceanic crust: an experimental study on H₂O–CO₂ bearing basalts. *Earth Planet. Sci. Lett.* **176**, 295–310 (2000).
5. Domanik, K. J. & Holloway, J. R. Experimental synthesis and phase relations of phengitic muscovite from 6.5 to 11 GPa in a calcareous metapelite from the Dabie Mountains, China. *Lithos* **52**, 51–77 (2000).
6. Caldeira, K. Continental-pelagic carbonate partitioning and the global carbonate-silicate cycle. *Geology* **19**, 204–206 (1991).

7. Kerrick, D. M. & Connolly, J. A. D. Metamorphic devolatilization of subducted oceanic metabasalts: Implications for seismicity, arc magmatism and volatile recycling. *Earth Planet. Sci. Lett.* (in the press).
8. Plank, T. & Langmuir, C. H. The chemical composition of subducting sediment and its consequences for the crust and mantle. *Chem. Geol.* **145**, 325–394 (1998).
9. Connolly, J. A. D. Multivariable phase diagrams: an algorithm based on generalized thermodynamics. *Am. J. Sci.* **290**, 666–718 (1990).
10. Holland, T. & Powell, R. An internally consistent thermodynamic data set for phases of petrologic interest. *J. Metamorph. Geol.* **16**, 309–343 (1998).
11. Holland, T. & Powell, R. A compensated Redlich-Kwong (CORK) equation for volumes and fugacities of CO₂ and H₂O in the range 1 bar to 50 kbar and 100–1600°C. *Contrib. Mineral. Petrol.* **109**, 265–273 (1991).
12. Peacock, S. M. & Wang, K. Seismic consequences of warm versus cool subduction metamorphism: Examples from Southwest and Northeast Japan. *Science* **286**, 937–939 (1999).
13. Becker, H. & Altherr, R. Evidence from ultra-high-pressure marbles for recycling of sediments into the mantle. *Nature* **358**, 745–748 (1992).
14. Schmidt, M. W. & Poli, S. Experimentally based water budgets for dehydrating slabs and consequences for arc magma generation. *Earth Planet. Sci. Lett.* **163**, 361–379 (1998).
15. Hacker, B. R. in *Subduction Top to Bottom* (eds Bebout, G. E., Scholl, D. W., Kirby, S. H. & Platt, J. P.) 337–346 (Monograph 96, American Geophysical Union, Washington DC, 1996).
16. Johnson, M. C. & Plank, T. Dehydration and melting experiments constrain the fate of subducted sediments. *Geochem. Geophys. Geosyst.* **1**, 1525–2027 (1999).
17. Kirby, S., Engdahl, E. R. & Denlinger, R. in *Subduction Top to Bottom* (eds Bebout, G. E., Scholl, D. W., Kirby, S. H. & Platt, J. P.) 195–214 (Monograph 96, American Geophysical Union, Washington DC, 1996).
18. Tatsumi, Y. & Eggins, S. *Subduction Zone Magmatism* (Blackwell Scientific, Oxford, 1995).
19. Alt, J. C. & Teagle, D. A. H. The uptake of CO₂ during alteration of the ocean crust. *Geochim. Cosmochim. Acta* **63**, 1527–1535 (1999).
20. Marty, B. & Tostikhin, I. N. CO₂ fluxes from mid-ocean ridges, arcs and plumes. *Chem. Geol.* **145**, 233–248 (1998).

Supplementary information is available on Nature's World-Wide Web Site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank S. Peacock for advice about subduction zone geotherms, and G. Bebout, K. Caldeira and M. Schmidt for comments and suggestions. This work was supported by the MARGINS program of the NSF.

Correspondence and requests for materials should be addressed to D.M.K. (e-mail: kerrick@geosc.psu.edu).

Adjustment to climate change is constrained by arrival date in a long-distance migrant bird

Christiaan Both* & Marcel E. Visser†

* Zoological Laboratory, Groningen University, PO Box 14, 9750 AA Haren, The Netherlands

† Netherlands Institute of Ecology, PO Box 40, 6666 ZG Heteren, The Netherlands

Spring temperatures in temperate regions have increased over the past 20 years¹, and many organisms have responded to this increase by advancing the date of their growth and reproduction^{2–7}. Here we show that adaptation to climate change in a long-distance migrant is constrained by the timing of its migratory journey. For long-distance migrants climate change may advance the phenology of their breeding areas, but the timing of some species' spring migration relies on endogenous rhythms that are not affected by climate change⁸. Thus, the spring migration of these species will not advance even though they need to arrive earlier on their breeding grounds to breed at the appropriate time. We show that the migratory pied flycatcher *Ficedula hypoleuca* has advanced its laying date over the past 20 years. This temporal shift has been insufficient, however, as indicated by increased selection for earlier breeding over the same period. The shift is hampered by its spring arrival date, which has not advanced. Some of the numerous long-distance migrants will suffer from

climate change, because either their migration strategy is unaffected by climate change, or the climate in breeding and wintering areas are changing at different speeds, preventing adequate adaptation.

Higher spring temperatures over the past two decades have led to advancing tree phenology and subsequently to earlier peaks in insect abundance^{9–11}. Several bird species^{3,4}, but not all^{9,12}, have advanced their egg-laying date as a consequence of this advancement in their food supply^{4–7,9}. Advancement of laying has not kept up with selection for earlier laying in areas where early spring temperatures have not increased as much as those late in spring^{9,13}. Adaptive advancement of reproduction may also be hampered in species that rely on endogenous rhythms or environmental stimuli unrelated to temperature⁸, such as day length. Long-distance migrants are extremely vulnerable in this sense, because in some well-studied species the timing of spring migration from the overwintering area is triggered by cues that are unlinked to the climate at their breeding grounds, and their breeding date is constrained by their arrival date.

We examined how a long-distance migrant bird, the pied flycatcher *Ficedula hypoleuca*, has responded to recent climate change, using data from a long-term study in the Hoge Veluwe (The Netherlands). Pied flycatchers overwinter in the zone of dry tropical forest at about 10° north in west Africa, and breed in temperate forests in Europe¹⁴. Males normally arrive before females, and females select males on the quality of their territory¹⁵.

Temperatures at the time of arrival and the start of breeding by pied flycatchers (16 April to 15 May) have increased significantly over the period 1980–2000 (Fig. 1a). Over the same period, the birds have not advanced the spring arrival on their breeding grounds (Fig. 1b), but have advanced their mean laying date by about 10 days (Fig. 1c). Mean laying date was unrelated to population size¹⁶, and population size did not show a significant trend ($r = -0.08$, $n = 21$, $P = 0.72$). Selection for early laying date has become stronger over the course of this 20-year period (Fig. 1d), indicating that early laying pairs do better than later pairs. Thus, the response

in laying date has not been sufficiently strong to track the advancement of spring.

Mean laying date was strongly correlated with the mean temperature in the second half of April and the first half of May (Fig. 2a), but arrival date was not (Fig. 2b). Arrival and laying dates were furthermore uncorrelated with temperatures in the first half of April, just before arrival of most birds (arrival: $r = -0.04$, $n = 20$, $P = 0.87$; laying: $r = 0.30$, $n = 21$, $P = 0.20$). The advancement of egg-laying date was, to a large extent, caused by plasticity of individual females to temperature¹² (analysis of covariance (ANCOVA): individual, $F_{272,411} = 2.16$, $P < 0.001$; age, $F_{1,412} = 17.12$, $P < 0.001$; temperature, $F_{1,412} = 50.57$, $P < 0.001$; average individual slope (-1.10 ± 0.15) significantly differs from population slope (-1.67 ± 0.24), $F_{1,413} = 13.55$, $P < 0.001$), and not by selection for genotypes for early breeding (heritability¹⁷ (h^2) for laying date (relative to yearly mean) calculated from mother-daughter regression is not significant: $F_{1,225} = 1.11$, $P = 0.29$, $h^2 = 0.16 \pm 0.15$). As a consequence of the correlation between laying and temperature (Fig. 2a) and the absence of a correlation between arrival date and temperature (Fig. 2b), the interval between arrival and egg-laying has decreased with increasing temperatures (linear regression: $F_{1,19} = 9.73$, $P = 0.006$).

Pied flycatchers were able to advance their laying date because they normally arrive on their breeding grounds earlier than their average optimal laying date¹⁴. Their spring migration strategy, triggered by day-length variation on their wintering grounds⁸, enabled them to arrive in time to respond adaptively to the naturally occurring variation in the start of spring¹⁸, and thereby start egg-laying at the date that maximizes fitness⁹. Owing to the advanced phenology on their breeding grounds and their relatively inflexible arrival date, however, this window has become too narrow, and a significant part of the population is now laying too late to exploit the peak in insect abundance optimally (as shown by the increasing selection for early laying; Fig. 1d).

The strong response of laying date in the pied flycatcher seems at variance with the lack of response in the resident great tit *Parus major* population breeding in the same area⁹. In both cases, however, the advancement of laying date is insufficient, as indicated by the presence of increased selection for earlier breeding. For the great tits, which start egg-laying about 2 weeks earlier than the flycatchers, the lack of a sufficient response is due to increasing temperatures in late, but not in early spring. For the flycatchers, the advancement is not hampered by such temporal variation in climate change, but because their timing of spring migration is triggered by day length, which is not affected by spring temperatures on their breeding grounds. The decision when to start spring migration thus becomes maladaptive if the cue used for migration is independent of the environmental change in the breeding area.

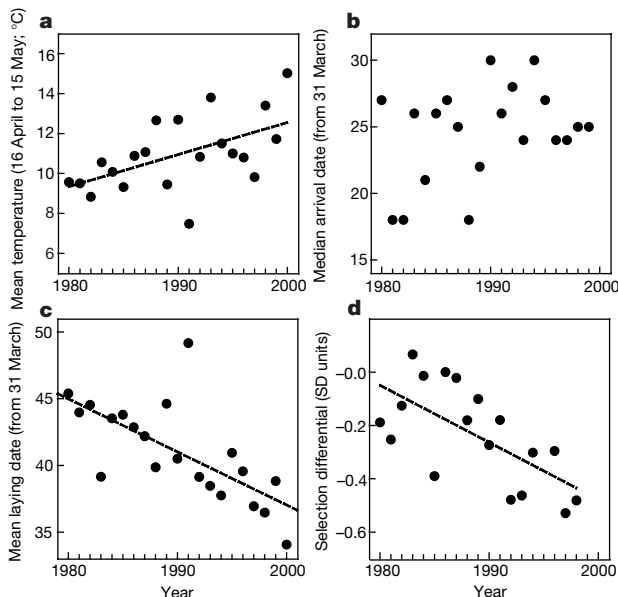


Figure 1 Spring temperature, breeding and spring arrival date of a pied flycatcher population in the Netherlands from 1980 to 2000. **a**, Mean average daily temperature in the period of arrival increased (linear regression: $F_{1,19} = 8.31$, $P < 0.001$). **b**, Median arrival date did not advance ($F_{1,19} = 2.32$, $P = 0.15$). **c**, Mean laying date did advance ($F_{1,20} = 17.14$, $P < 0.001$). **d**, Standardized selection differential for laying date decreased with year ($F_{1,17} = 12.00$, $P = 0.003$) over this period.

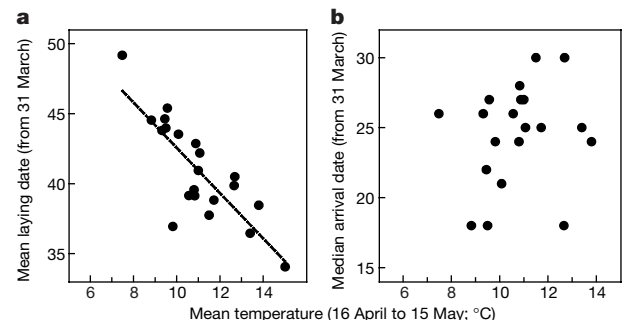


Figure 2 Breeding date and spring arrival of a pied flycatcher population as a function of spring temperature in the Netherlands. **a**, Average laying date was significantly related to the mean average daily temperature (°C) in the second half of April and the first half of May ($F_{1,20} = 37.59$, $P < 0.001$). **b**, Median arrival date was unrelated to spring temperature ($F_{1,19} = 0.66$, $P = 0.43$).

In other long-distance migrants, arrival on the breeding grounds is also relatively insensitive to the temperature on arrival¹⁹, although in some species the temperature on journey is a good correlate of arrival date^{20,21}. Climatic factors may be important in fine-tuning the onset and speed of migration, but climate change differs between temperate and tropical latitudes¹, and therefore a response to environmental cues such as temperature for the onset and speed of migration may not lead to an adequate arrival date on the breeding grounds. Short-distance migrants may be more flexible in their response, because the circumstances on the wintering grounds will be a better predictor for the optimal arrival time on the breeding grounds and genetic variation has been shown for some of their migratory traits^{22,23}.

Large-scale climate change may thus form a serious threat to at least some of the numerous species that migrate from tropical wintering grounds to temperate breeding areas²⁴, because they arrive at an inappropriate time to exploit the habitat optimally, and face higher competition with resident species that may have increased in numbers through enhanced winter survival²⁵. This may, in fact, be partly responsible for the decline of these species in western Europe²⁶. □

Methods

Data collection

Data were collected from a nest-box breeding population of pied flycatchers in the Hoge Veluwe area, central Netherlands, between 1965 and 2000 (ref. 9). We analysed data from 1980 to 2000, because temperature increased most markedly after 1980. We used only nests that were considered to be first nesting attempts of females during that year ($n = 1,892$). Parents and chicks were ringed with uniquely numbered aluminium rings. Arrival data were obtained from a local amateur bird group, working within 10 km from the study area²⁷. Members of this bird group recorded each year the first singing pied flycatcher, and the median first arrival date was used as approximation of arrival in the study area. The lack of an advancement in arrival date was confirmed by analysing the arrival date of the first male recorded in the study area from 1992 to 2000 ($F_{1,6} = 0.26$, $P = 0.63$), the first ten males that arrived in an area nearby from 1980 to 1990 ($F_{1,10} = 0.02$, $P = 0.88$), and the mean start of nest building of the first ten pied flycatcher nests each year in 1980–2000 ($F_{1,19} = 0.92$, $P = 0.35$; start of egg-laying was estimated from the state of the nest during weekly checks).

Analyses

All analyses were performed with linear regression using two-tailed P values. In most cases we used annual means. In the analysis of the response of individual females to temperature, we used females that bred in at least 2 years ($n = 273$). In this analysis, female age was a factor for first known breeding or later breeding in the area (real age was not determined), because first year breeders normally breed later¹⁴. Individual is used here as a factor in an ANCOVA. Temperature used is the average of the mean daily temperatures from 16 April to 15 May recorded by the Royal Dutch Meteorological Institute (KNMI) at De Bilt (The Netherlands). The standardized selection differential is the mean laying date weighted for the number of recruits (offspring that return as breeding birds in the study area) each nest produced minus the mean laying date, divided by the standard deviation of laying date¹⁷. Selection differentials are given until 1998, because the number of recruits for later years is not yet known.

Received 1 November 2000; accepted 28 February 2001.

- Houghton, J. T. et al. (eds) *Climate Change 1995* (Cambridge Univ. Press, Cambridge, 1996).
- Beebe, T. J. C. Amphibian breeding and climate. *Nature* **374**, 219–220 (1995).
- Crick, H. Q. P., Dudley, C., Glue, D. E. & Thomson, D. L. UK birds are laying eggs earlier. *Nature* **388**, 526 (1997).
- Crick, H. Q. P. & Sparks, T. H. Climate change related to egg-laying trends. *Nature* **399**, 423–424 (1999).
- Forchhammer, M. C., Post, E. & Stenseth, N. C. Breeding phenology and climate. *Nature* **391**, 29–30 (1998).
- McClery, R. H. & Perrins, C. M. ...temperature and egg-laying trends. *Nature* **391**, 30–31 (1998).
- Brown, J. L., Li, S. H. & Bhagabati, N. Long-term trend toward earlier breeding in an American bird: a response to global warming? *Proc. Natl Acad. Sci. USA* **96**, 5565–5569 (1999).
- Gwinner, E. Circannual clocks in avian reproduction and migration. *Ibis* **138**, 47–63 (1996).
- Visser, M. E., van Noordwijk, A. J., Tinbergen, J. M. & Lessells, C. M. Warmer springs lead to mistimed reproduction in great tits (*Parus major*). *Proc. R. Soc. Lond. B* **265**, 1867–1870 (1998).
- Harrington, R., Woiwod, I. & Sparks, T. H. Climate change and trophic interactions. *Trends Ecol. Evol.* **14**, 146–150 (1999).
- Visser, M. E. & Holleman, L. J. M. Warmer spring disrupt the synchrony of Oak and Winter Moth phenology. *Proc. R. Soc. Lond. B* **268**, 289–294 (2001).
- Przybylo, R., Sheldon, B. C. & Merila, J. Climatic effect on breeding and morphology: evidence for phenotypic plasticity. *J. Anim. Ecol.* **69**, 395–403 (2000).
- Stevenson, I. R. & Bryant, D. M. Climate change and constraints on breeding. *Nature* **406**, 366–367 (2000).

- Lundberg, A. & Alatalo, R. V. *The Pied Flycatcher* (T. & A. D. Poyser, London, 1992).
- Alatalo, R. V., Lundberg, A. & Glynn, C. Female Pied Flycatchers choose territory quality and not male characteristics. *Nature* **323**, 152–153 (1986).
- Both, C. Density dependence of avian clutch size in resident and migrant species: is there a constraint on the predictability of competitor density? *J. Avian Biol.* **31**, 412–417 (2000).
- Falconer, D. S. & Mackay, E. T. C. *An Introduction to Quantitative Genetics* 4th edn (Longman Higher Education, Essex, 1996).
- Buse, A., Dury, S. J., Woodburn, R. J. W., Perrins, C. M. & Good, J. E. G. Effects of elevated temperature on multi-species interactions: the case of Pedunculate Oak, Winter Moth and Tits. *Funct. Ecol.* **13** (suppl. 1), 74–82 (1999).
- Mason, C. F. Long term trends in the arrival dates of spring migrants. *Bird Study* **42**, 182–189 (1995).
- Huin, N. & Sparks, T. H. Spring arrival patterns of the Cuckoo *Cuculus canorus*, Nightingale *Luscinia megarhynchos* and Spotted Flycatcher *Muscicapa striata* in Britain. *Bird Study* **47**, 22–31 (2000).
- Huin, N. & Sparks, T. H. Arrival and progression of the Swallow *Hirundo rustica* through Britain. *Bird Study* **45**, 361–370 (1998).
- Berthold, P., Helbig, A. J., Mohr, G. & Querner, U. Rapid microevolution of migratory behaviour in a wild bird species. *Nature* **360**, 668–670 (1992).
- Pulido, F., Berthold, P. & van Noordwijk, A. J. Frequency of migrants and migratory activity are genetically correlated in a bird population: evolutionary implications. *Proc. Natl Acad. Sci. USA* **93**, 14642–14647 (1996).
- Berthold, P., Fiedler, W., Schlenker, R. & Querner, U. 25-year study of the population development of Central European songbirds: A general decline, most evident in long-distance migrants. *Naturwissenschaften* **85**, 350–353 (1998).
- Berthold, P. Patterns of avian migration in light of current global 'greenhouse' effects: a central European perspective. *Acta XX Congressus Internationalis Ornithologici* 780–786 (1990).
- Myers, J. P. & Lester, R. T. *Global Warming and Biodiversity* (eds Peters, R. L. & Lovejoy, T. E.) 193–200 (Yale Univ. Press, New Haven, 1992).
- Pilzecker, J. Een voorjaar maakt nog geen fenologie; Bewerking van de fenologiegegevens van Vogelwerkgroep Arnhem e. o. over de periode 1980–1997. Report No. 98-2 (Vogelwerkgroep Arnhem e. o., Arnhem, 1998).

Acknowledgements

J. H. van Balen and J. Visser managed the long-term study for many years. The board of National Park 'De Hoge Veluwe' kindly allowed us to work on their property. J. Pilzecker and R. Bijlsma allowed us to use some of their unpublished observations of flycatcher arrival. J. Harvey, C. M. Lessells and A. J. van Noordwijk commented on an earlier draft.

Correspondence and requests for materials should be addressed to M.E.V. (e-mail: m.visser@cto.nioo.knaw.nl).

Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm

Nicole Dubilier*, Caroline Mülders*, Tim Ferdelman*, Dirk de Beer*, Annelie Pernthaler*, Michael Klein†, Michael Wagner†, Christer Erséus‡, Frank Thiermann§, Jens Krieger§, Olav Giere§ & Rudolf Amann*

* Max Planck Institute of Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany

† Department of Microbiology, Technical University of Munich, D-85350 Freising, Germany

‡ Department of Invertebrate Zoology, Swedish Museum of Natural History, S-10405 Stockholm, Sweden

§ Zoological Institute and Zoological Museum, University of Hamburg, D-20146 Hamburg, Germany

Stable associations of more than one species of symbiont within a single host cell or tissue are assumed to be rare in metazoans because competition for space and resources between symbionts can be detrimental to the host¹. In animals with multiple endosymbionts, such as mussels from deep-sea hydrothermal vents² and reef-building corals³, the costs of competition between the symbionts are outweighed by the ecological and physiological flexibility gained by the hosts. A further option for the coexistence of multiple symbionts within a host is if these benefit directly from one another, but such symbioses have not been previously described. Here we show that in the gutless marine oligochaete

Olavius algarvensis, endosymbiotic sulphate-reducing bacteria produce sulphide that can serve as an energy source for sulphide-oxidizing symbionts of the host. Thus, these symbionts do not compete for resources but rather share a mutualistic relationship with each other in an endosymbiotic sulphur cycle, in addition to their symbiotic relationship with the oligochaete host.

*Olavius algarvensis*⁴ is a small tubificid worm (0.2 mm × 20–30 mm) that is found in the Mediterranean at sediment depths of 5–15 cm in coarse-grained sands surrounding beds of sea grass. As in other gutless oligochaetes^{5,6}, two bacterial morphotypes occur in immediate proximity to one another just below the cuticle between extensions of the epidermal cells (Fig. 1). The larger morphotype (2.5 µm × 1.5 µm) contains numerous intracellular globules, whereas the smaller (1.1 µm × 0.7 µm) has no conspicuous inclusions.

We determined the phylogenetic identity of the *O. algarvensis* symbionts by using comparative 16S ribosomal RNA sequencing. We identified two dominant clone groups in the hosts, with minimal variations in the 16S rRNA sequences within each clone group (0.1–1.2%). Phylogenetic analyses revealed that the 16S rRNA sequences from these two groups are derived from the γ- and δ-subclasses of the Proteobacteria (Fig. 2a, b). The γ-proteobacterial sequence isolated from *O. algarvensis* consistently falls in a cluster with endosymbionts from other gutless oligochaetes such as *Olavius loisae*⁷ and *Inanidrillus leukodermatus*⁸ (96–97% sequence identity) in all treeing methods used. The δ-proteobacterial sequence is always placed within a subgroup of free-living sulphate-reducing bacteria (*Desulfococcus*/*Desulfonema*/*Desulfosarcina*) by all inference methods, with *Desulfosarcina variabilis* consistently identified as its closest relative (93% sequence identity).

Fluorescence *in situ* hybridization (FISH) confirmed that the γ- and δ-proteobacterial 16S rRNA sequences originated from the symbiotic bacteria in *O. algarvensis* (Fig. 3). The FISH signal from the probe specific to the γ-subclass of the Proteobacteria (GAM42a) and a species-specific probe based on the *O. algarvensis* γ-sequence (OalgGAM445) clearly originated from the larger bacterial symbiont, whereas the general *Desulfosarcina*/*Desulfococcus* probe (DSS658) and a probe targeting the *O. algarvensis* δ-sequence

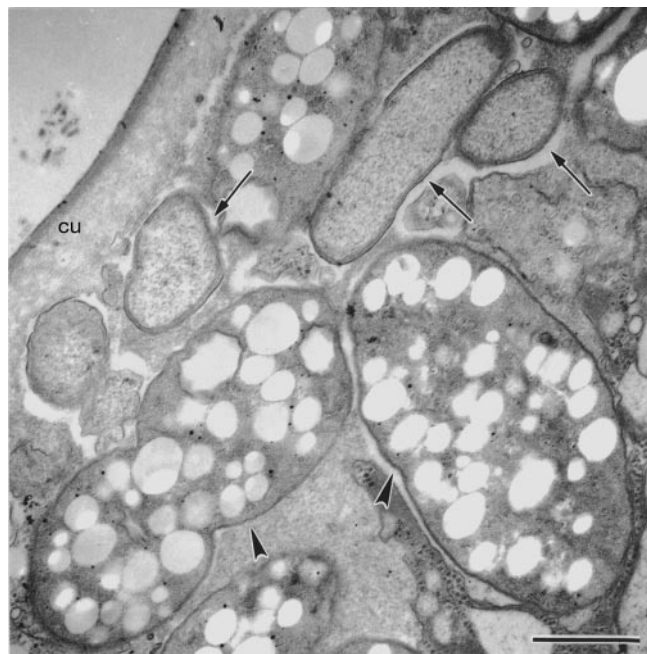


Figure 1 Transmission electron micrograph of bacterial endosymbionts in *O. algarvensis*. The symbionts occur just below the cuticle (cu) between extensions of the epidermal cells. The larger bacteria (arrowheads) contain numerous globules whereas the smaller bacteria (arrows) do not show any cytoplasmic inclusions. Scale bar, 1 µm.

(OalgDEL136) consistently labelled the smaller bacterial symbiont.

The thioautotrophic nature (that is, sulphur-oxidizing, CO₂-fixing metabolism) of the γ-symbionts in *O. algarvensis* is suggested by their close evolutionary relationship to symbionts already characterized as thioautotrophic^{8,9}. This assumption is corroborated by our results from immunocytochemical labelling with an antiserum directed against form I of ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco), the key CO₂-fixing enzyme. The antiserum consistently labelled the larger γ-symbionts but not the smaller δ-symbionts (see Supplementary Information). Further evidence for a thioautotrophic metabolism of the γ-symbionts is

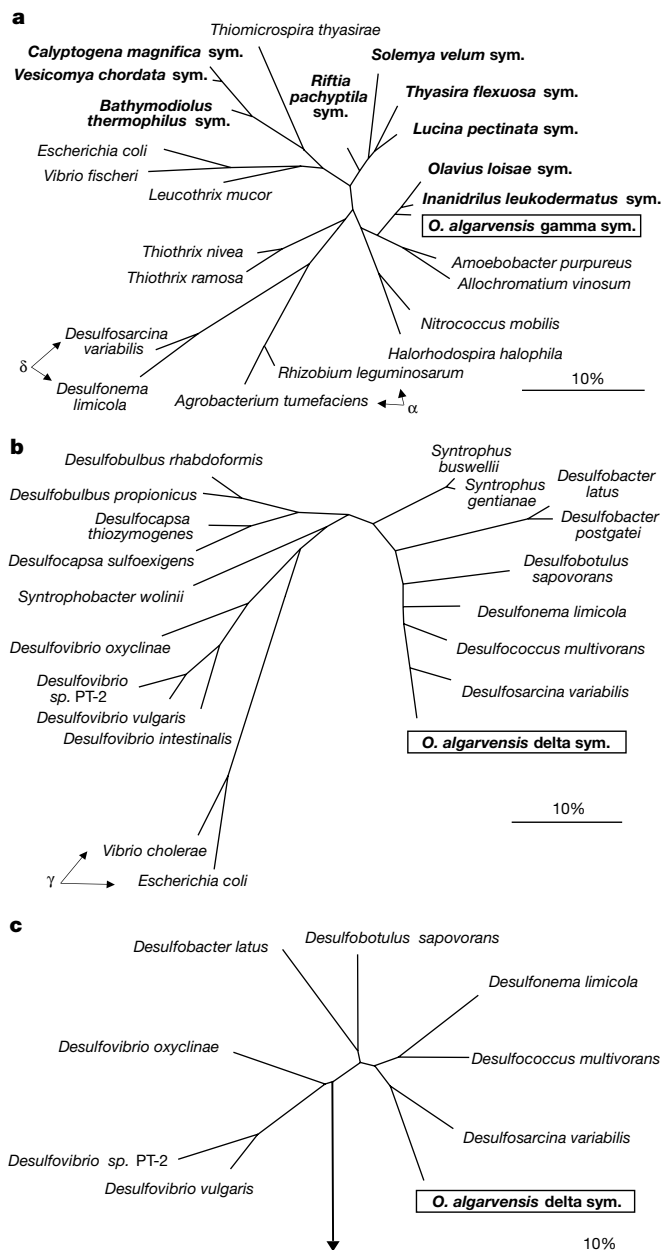


Figure 2 Phylogenetic relationships of the *O. algarvensis* symbionts based on maximum likelihood analyses. **a**, 16S rRNA sequences from the γ-subclass of Proteobacteria (α- and δ-Proteobacteria marked with arrows; chemoautotrophic symbionts (sym.) in bold type). **b**, 16S rRNA sequences from the δ-subclass of Proteobacteria (γ-Proteobacteria marked with arrows). **c**, DSR sequences based on a concatenated amino-acid alignment encompassing the DSR α- and β-subunit data sets. Arrow indicates published¹⁶ and unpublished DSR sequences (M.W. *et al.*) not shown in tree. Scale bars indicate 0.10 expected substitutions per site.

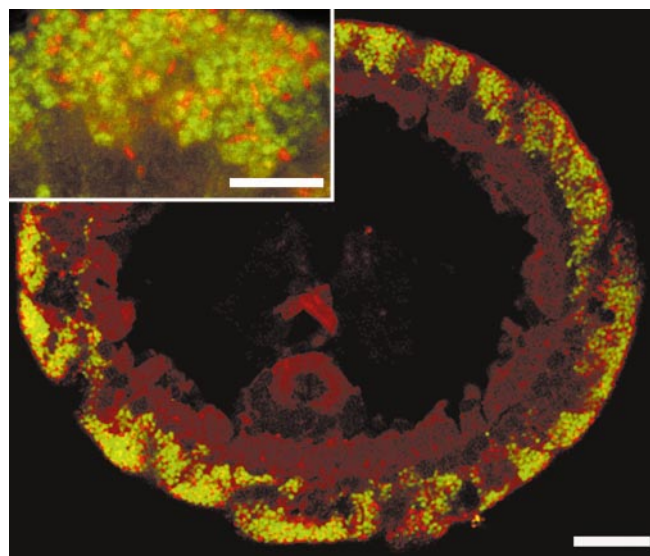


Figure 3 Fluorescence *in situ* hybridization of endosymbionts in *O. algarvensis* with oligonucleotide probes labelled with fluorochromes. γ -symbionts are green, δ -symbionts are red. Cross-section through entire worm. Hybridization with probes for the γ -subclass of the Proteobacteria and the *Desulfosarcina/Desulfococcus* group. Scale bar, 20 μ m. Inset: body wall of the worms with symbiont-containing region. Hybridization with specific probes for the γ - and δ -symbionts. Scale bar, 10 μ m.

the high concentration of elemental sulphur in *O. algarvensis* ($3.2 \pm 1.7\%$ dry weight; $n = 5$). Such large amounts of S^0 are characteristic for hosts with sulphide-oxidizing symbionts¹⁰. This corresponds well with electron microscopic spectroscopy studies that show the presence of sulphur in globules of the γ -symbionts (J.K., unpublished results).

The close evolutionary relationship of the δ -symbionts of *O. algarvensis* to free-living sulphate-reducing bacteria (SRB) suggests that these are also sulphate reducers. SRB have been described from termite guts¹¹ and the intestines of some mammals¹² and there is indirect evidence that they may occur as epibionts on some marine ciliates¹³ and invertebrates^{14,15}. However, SRB as endosymbionts have not been previously found in marine invertebrates and it has been suggested that such symbioses are unlikely because sulphide, their metabolic endproduct, is toxic to most aerobic organisms. We therefore used several methods to show that the δ -symbionts of *O. algarvensis* are indeed SRB and can actively respire sulphate in the worms.

The enzyme dissimilatory sulphite reductase (DSR) catalyses the reduction of (bi)sulphite to sulphide and is a good indicator for dissimilatory sulphate respiration, as it is only known to occur in sulphate-reducing prokaryotes¹⁶. Using specific primers, we successfully amplified the gene encoding DSR from *O. algarvensis*; no

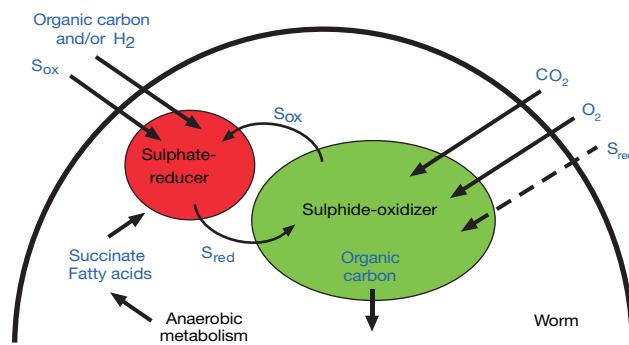


Figure 4 Model of the endosymbiotic sulphur cycle in *O. algarvensis* showing syntrophic cycling of oxidized and reduced sulphur compounds between the sulphate-reducing and sulphide-oxidizing symbionts. Under typical conditions of low sulphide flux from the sediment, sulphide produced internally by the sulphate reducers is used by the sulphide oxidizers as an electron donor for the autotrophic fixation of CO_2 . Electron donors such as succinate and fatty acids can be supplied to the sulphate reducer internally by the worm during anaerobic metabolism. For net growth of the symbiotic association, external electron donors (organic carbon or H_2) are taken up from the sediment.

amplification products were obtained from negative controls with another gutless oligochaete host (*I. leukodermatus*) that does not harbour δ -proteobacterial symbionts. Comparative phylogenetic analyses (Fig. 2c) consistently showed that the DSR sequence from *O. algarvensis* is most closely related to *D. variabilis* (79% DNA sequence identity, 82% amino-acid identity), the free-living SRB most closely related to the δ -symbiont of *O. algarvensis* on the basis of 16S rRNA analyses (Fig. 2b). Previous studies have shown that 16S rRNA phylogenies of SRB agree well with their DSR phylogenies¹⁶, indicating that the DSR sequence isolated from *O. algarvensis* originated from the δ -symbiont of this host and thus that this symbiont is a sulphate reducer.

To show that sulphate is actively reduced in *O. algarvensis*, we inserted silver needles through individual worms and incubated these in radiolabelled $^{35}SO_4^{2-}$ under microaerobic and aerobic conditions. After exposure of the needles to an autoradiographic film, blots from the needles inserted in live worms under microaerobic conditions showed a positive signal from ^{35}S -labelled sulphide that had precipitated on the needles, whereas under the same conditions a needle inserted in a formalin-fixed worm remained unlabelled (data not shown). This indicates that sulphate is reduced to sulphide during dissimilatory sulphate respiration by the δ -symbionts of *O. algarvensis* under microaerobic conditions. Sulphate respiration appears to be inhibited at high O_2 concentrations, on the basis of the absence of a sulphide precipitate on needles inserted in worms incubated under aerobic conditions.

We determined the sulphate reduction rates (SRRs) of the symbionts by incubating *O. algarvensis* in $^{35}SO_4^{2-}$ under microaerobic conditions (Table 1). In live worms, we measured SRRs of 53–

Table 1 Sulphate reduction rates in *O. algarvensis* specimens

Substrate	Worms	No. rep*	O_2 (μ M)†	Sulphate reduction rate (pmol per worm per day)‡	Sulphate reduction rate (nmol cm^{-3} per day)§
Agar	Live	4	2–4	115 ± 76 (53/72/113/223)	$1,474 \pm 957$ (690/935/1,440/2,830)
Agar	Live	1	200	<7.2	<80
Agar	Dead	1	2–4	<7.2	<80
Sand	Live	3	Micro-aerobic	638 ± 779 (115/266/1,534)	$8,213 \pm 9,917$ (1,470/3,570/19,600)
Sand	Live	1	Aerobic	<7.2	<80
Sand	Dead	1	Micro-aerobic	<7.2	<80

Specimens were incubated in agar or sterilized sand.

*Number of replicate experiments; values in parentheses are rates measured in each replicate experiment.

† Measured at oligochaete surface in agar experiments (oxygen conditions in sand experiments are estimated from the agar experiments, as an identical experimental setup was used).

‡ Detection limit, 7.2 pmol per worm per day.

§ Detection limit: 80 nmol cm^{-3} per day.

1,534 pmol per worm per day, whereas SRRs in heat-killed worms under the same conditions were below detection limits. Sulphate was reduced to sulphide despite the absence of an external electron donor in the incubation medium. Endogenous electron donors that could have been used by the sulphate-reducing symbionts are fermentation products from the host such as succinate, propionate and acetate. These substrates accumulate during anaerobic metabolism under low oxygen concentrations in other marine tubificids¹⁷ and many other aquatic invertebrates¹⁸. Under fully aerobic conditions, SRRs in live worms were below detection limits, indicating, as in the silver needle experiments, that high oxygen concentrations inhibit sulphate reduction. This corresponds well with observations on SRB in pure cultures, where most species are temporarily oxygen tolerant but not able to respire sulphate in the presence of high oxygen concentrations¹⁹. On the basis of the numbers of δ -symbionts in *O. algarvensis* as estimated by FISH, SRRs in these hosts (0.07–0.36 fmol per cell per day) are lower than those of SRB in pure cultures with saturating substrate concentrations (0.2–50 fmol per cell per day)²⁰ but in the same range as those estimated for free-living SRB in marine sediments (0.01–0.09 fmol per cell per day)²¹. SRRs in the worms on a volumetric basis are extremely high (690–19,600 nmol cm⁻³ per day) and comparable with rates measured in microbial mats (2,880–43,200 nmol cm⁻³ per day)²².

To estimate the importance to the sulphide-oxidizing symbionts of internally produced sulphide compared with the import of external sulphide from the sediment, we compared the fluxes from these two sulphide sources. Dissolved sulphide concentrations in pore waters of *O. algarvensis* collection sites were extremely low: <14–76 nM (26 ± 21 , $n = 9$) at 5–15 cm sediment depth, with no trend with sediment depth or location. Correspondingly, sulphide flux from the environment into the worm was <50–270 pmol per worm per day (93 ± 75 , $n = 9$). Internal sulphide production from the sulphate-reducing symbionts on the basis of SRRs was 120–1,530 pmol per worm per day (640 ± 780 , $n = 3$). Thus, internal sulphide production is typically considerably higher than sulphide flux from the sediment, indicating that under prevalent conditions this symbiosis appears to be independent of an external source of sulphide.

The coexistence of sulphate-reducing and sulphide-oxidizing bacteria as endosymbionts in *O. algarvensis* indicates that these are engaged in a syntrophic sulphur cycle in which oxidized and reduced sulphur compounds are recycled between the two symbionts (Fig. 4). For net growth of the symbiotic association, uptake of organic or inorganic sources of carbon and electron donors from the environment is required. As sulphide flux calculations indicate that the electron donor for the sulphide oxidizers is typically supplied internally, external reductants must be imported through the sulphate reducers. Given the metabolic diversity of SRB, in particular within the *Desulfosarcina* group, where both chemo-organotrophic and chemoautotrophic metabolism occurs, dissolved organic carbon and hydrogen are possible sources of reducing power. Migration of the worms between oxidized and reduced sediments, as described for other gutless oligochaetes²³, would provide the host and its sulphide-oxidizing symbionts with oxygen and the sulphate reducers with reductants. The benefits of this endosymbiotic sulphur cycle to its partners are clear. Cycling of oxidized and reduced sulphur compounds between the two symbionts would result in increased protein yields, as shown for continuous cultures with free-living SRB and sulphide-oxidizing bacteria²⁴. Furthermore, fermentation products of the host that accumulate during anaerobic metabolism would provide the sulphate reducers with an ideal energy source, aid the hosts in the removal of these undesirable endproducts and recycle metabolites that would otherwise be lost to the symbiosis. A further advantage for the host and its thioautotrophic symbiont is that they are not limited by the external presence of reduced sulphur compounds, given the endogenous production of sulphide by the sulphate-

reducing symbiont. Thus the uptake of a sulphate reducer may have enabled these hosts to colonize new habitats and extend their geographic distribution. □

Methods

For more details see Supplementary Information.

Specimens

O. algarvensis was collected in 1998–2000 from sediments at 6–8 m water depth in a bay off Capo di San Andrea (Elba, Italy) by SCUBA divers. *I. leukodermatus* specimens used as negative controls for the DSR amplifications were collected in Bermuda in 1997.

Pore water sulphide

Pore water was collected at 5, 10 and 15 cm depth at the *O. algarvensis* collection sites by SCUBA divers with immediate fixation of the samples in zinc acetate. In June 1999, October 1999 and January 2000, 1–2 ml of pore water per sample was collected and total sulphide concentrations were below the detection limit of 0.4 μ M in all samples. In June 2000 the detection limit was lowered to 14 nM by collecting greater amounts of pore water (40–60 ml per sampling site) using samplers connected to evacuated serum vials containing zinc chloride. Concentrations of total sulphide were determined colorimetrically²⁵.

Transmission electron microscopy and immunocytochemistry

O. algarvensis individuals were fixed and prepared for electron microscopy as described⁴. For Rubisco immunocytochemistry, specimens were treated as described in ref. 9. In each worm ($n = 5$) 50–100 symbionts were examined for labelling response.

DNA analyses

Three *O. algarvensis* individuals (and two *I. leukodermatus* specimens for DSR controls) were prepared singly for polymerase chain reaction (PCR) as described in ref. 7. DNA was isolated from *D. variabilis* DSM 2060 as described¹⁶. Amplifications were performed with primers specific for the bacterial 16S rRNA genes (8F and 1507R) or the DSR genes of SRB (DSR1F and DSR4R)¹⁶. PCR products were cloned and grouped using amplified ribosomal DNA restriction analysis (ARDRA). Two or three clones per individual from dominant ARDRA groups were partially sequenced and at least one clone per individual from each ARDRA group was sequenced fully in both directions. Alignments, treeing and phylogenetic analyses (distance, parsimony and maximum likelihood) were performed with the ARB program (<http://www.mikro.biologie.tu-muenchen.de/pub/ARB/>).

FISH

Five worms were fixed and prepared for FISH as described⁷. Sections were hybridized as described⁷ with Cy3 and Cy5 labelled group-specific probes (GAM42a and DSS658) as well as two specific probes designed for this study (OalgGAM445: 5'-CTCGAGATCTTTCTTCCC-3'; OalgDEL136: 5'-GTTATCCCGACTCGGGG-3'). Specificity of the probes was tested with reference strains as described⁷.

³⁵SO₄²⁻ incubations

For silver needle experiments worms were incubated in Na³⁵SO₄²⁻ and 0.2- μ m pore-size filtered seawater from the collection site. The medium was solidified with agar and the worms paralysed with lidocaine (2 mg ml⁻¹) to prevent excessive movements during insertion with silver needles (99.999% pure 50 μ M Ag wire, tapered to a <1 μ m tip). Incubations were run for 2–3 h under microaerobic (2–4 μ M O₂) and aerobic (200 μ M O₂) conditions with monitoring of oxygen concentrations with microsensors (two replicate experiments per O₂ concentration with one worm per incubation). In a control experiment at 2–4 μ M O₂ with a dead worm, the specimen was fixed in 4% formalin in seawater and subsequently washed in filtered seawater. After removal, the needles were washed in 50 mM Na₂SO₄ solution and exposed to autoradiography film. Results were similar in replicate experiments.

For determination of SRRs we incubated five worms per experiment for 2–3 h in seawater with Na³⁵SO₄²⁻ using agar or sand as substrates. Sand incubations were prepared and run in the same manner as the agar experiments (see above), but worms were not paralysed and moved freely in sand from the collection site that had been washed and combusted at 480 °C. Oxygen concentrations were not monitored during the sand incubations. For control experiments, specimens were heat killed in water at 70 °C for 10 min. SRRs were determined using the one-step acidic Cr-II method to separate reduced ³⁵S (ref. 26).

Elemental sulphur analyses

S⁰ was extracted individually from five worms with methanol and quantified by high-performance liquid chromatography as described²⁷.

Flux calculations

Sulphide flux (Q) from the environment was calculated using the following equation²⁸: $Q = 2\pi l D_{\text{eff}} C_p / \ln(1 + 2\delta/d)$, where the length of the worm (l) is 1 cm, the effective diffusion coefficient of total sulphide in sediment (D_{eff}) is 1.39×10^{-9} m² s⁻¹, C_p the concentration of total sulphide in the pore water, the mass boundary layer (δ) is 100 μ m and the diameter

(d) of the worm is 200 μm (see Supplementary Information). All assumptions are conservative and result in an overestimation of sulphide flux from the sediment (see Supplementary Information). Internal sulphide production from the symbionts is based on SRRs measured in worms incubated in sand (Table 1), assuming that all sulphide produced is consumed by the sulphide-oxidizing symbionts. SRRs in the worms are assumed to be underestimated, given that no external electron donor was used and experimental conditions are suboptimal in comparison to the natural environment.

Received 25 September 2000; accepted 26 February 2001.

- Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (Oxford Univ. Press, Oxford 1995).
- Distel, D. L., Lee, H. K.-W. & Cavanaugh, C. M. Intracellular coexistence of methano- and thioautotrophic bacteria in a hydrothermal vent mussel. *Proc. Natl Acad. Sci. USA* **92**, 9598–9602 (1995).
- Rowan, R., Knowlton, N., Baker, A. & Jara, J. Landscape ecology of algal symbionts creates variation in episodes of coral bleaching. *Nature* **388**, 265–266 (1997).
- Giere, O., Erséus, C. & Stühlmacher, F. A new species of *Olavius* (Tubificidae, Phalloporilinae) from the Algarve Coast in Portugal, the first East Atlantic gutless oligochaete with symbiotic bacteria. *Zool. Anzeiger* **237**, 209–214 (1998).
- Giere, O. & Langheld, C. Structural organisation, transfer and biological fate of endosymbiotic bacteria in gutless oligochaetes. *Mar. Biol.* **93**, 641–650 (1987).
- Giere, O., Nieser, C., Windoffer, R. & Erséus, C. A comparative structural study on bacterial symbioses of Caribbean gutless Tubificidae (Annelida, Oligochaeta). *Acta Zool.* **76**, 281–290 (1995).
- Dubilier, N. *et al.* Phylogenetic diversity of bacterial endosymbionts in the gutless marine oligochaete *Olavius loisei* (Annelida). *Mar. Ecol. Prog. Ser.* **178**, 271–280 (1999).
- Dubilier, N., Giere, O., Distel, D. L. & Cavanaugh, C. M. Characterization of chemotrophic bacterial symbionts in a gutless marine worm (Oligochaeta, Annelida) by phylogenetic 16S rRNA sequence analysis and *in situ* hybridization. *Appl. Environ. Microbiol.* **61**, 2346–2350 (1995).
- Krieger, J., Giere, O. & Dubilier, N. Localization of RubisCO and sulfur in endosymbiotic bacteria of the gutless marine oligochaete *Inanidrilus leukodermatus* (Annelida). *Mar. Biol.* **137**, 239–244 (2000).
- Vetter, R. D. & Fry, B. Sulfur contents and sulfur-isotope compositions of thiotrophic symbioses in bivalve molluscs and vestimentiferan worms. *Mar. Biol.* **132**, 453–460 (1998).
- Kühnigk, T., Branke, J., Krekeler, D., Cypionka, H. & König, H. A feasible role of sulfate-reducing bacteria in the termite gut. *System. Appl. Microbiol.* **19**, 139–149 (1996).
- Morvan, B., Bonnemoy, F., Fonty, G. & Gouet, P. Quantitative determination of H_2 -utilizing acetogenic and sulfate-reducing bacteria and methanogenic archaea from digestive tracts of different mammals. *Curr. Microbiol.* **32**, 129–133 (1996).
- Fenchel, T. & Ramsing, N. B. Identification of sulfate-reducing ectosymbiotic bacteria from anaerobic ciliates using 16S rRNA binding oligonucleotide probes. *Arch. Microbiol.* **158**, 394–397 (1992).
- Bussmann, I. & Reichardt, W. Sulfate-reducing bacteria in temporarily toxic sediments with bivalves. *Mar. Ecol. Prog. Ser.* **78**, 97–102 (1991).
- Cottrell, M. T. & Cary, C. S. Diversity of dissimilatory bisulfite reductase genes of bacteria associated with the deep-sea hydrothermal vent polychaete *Alvinella pompejana*. *Appl. Environ. Microbiol.* **65**, 1127–1132 (1999).
- Wagner, M., Roger, A. J., Flax, J. L., Brusseau, G. A. & Stahl, D. A. Phylogeny of dissimilatory sulfite reductases supports an early origin of sulfate respiration. *J. Bacteriol.* **180**, 2975–2982 (1998).
- Dubilier, N., Giere, O. & Grieshaber, M. K. Morphological and ecophysiological adaptations of the marine oligochaete *Tubificoides benedii* to sulfidic sediments. *Am. Zool.* **35**, 163–173 (1995).
- Grieshaber, M. K., Hardewig, I., Kreutzer, U. & Pörtner, H.-O. Physiological and metabolic responses to hypoxia in invertebrates. *Rev. Physiol. Biochem. Pharmacol.* **125**, 43–147 (1994).
- Barton, L. L. *Sulfate-Reducing Bacteria* (Plenum, New York, 1995).
- Jørgensen, B. B. A comparison of methods for the quantification of bacterial sulfate reduction in coastal marine sediments. III. Estimation from chemical and bacteriological field data. *Geomicrobiol. J.* **1**, 49–64 (1978).
- Sahm, K., MacGregor, B. J., Jørgensen, B. B. & Stahl, D. A. Sulfate reduction and vertical distribution of sulfate-reducing bacteria quantified by rRNA slot-blot hybridization in a coastal marine sediment. *Environ. Microbiol.* **1**, 65–74 (1999).
- Canfield, D. E. & Des Marais, D. J. Biogeochemical cycles of carbon, sulfur, and free oxygen in a microbial mat. *Geochim. Cosmochim. Acta* **57**, 3971–3984 (1993).
- Giere, O., Conway, N. M., Gastrock, G. & Schmidt, C. 'Regulation' of gutless annelid ecology by endosymbiotic bacteria. *Mar. Ecol. Prog. Ser.* **68**, 287–299 (1991).
- van den Ende, F. P., Meier, J. & van Gemerden, H. Syntrophic growth of sulfate-reducing bacteria and colorless sulfur bacteria during oxygen limitation. *FEMS Microbiol. Ecol.* **23**, 65–80 (1997).
- Cline, J. D. Spectrophotometric determination of hydrogen sulfide in natural waters. *Limnol. Oceanogr.* **14**, 454–458 (1969).
- Fossing, H. & Jørgensen, B. B. Measurements of bacterial sulfate reduction in sediments: Evaluation of a single-step chromium reduction method. *Biogeochemistry* **8**, 205–222 (1989).
- Ferdeman, T. G. *et al.* Sulfate reduction and methanogenesis in a *Thioploca*-dominated sediment off the coast of Chile. *Geochim. Cosmochim. Acta* **61**, 3065–3079 (1997).
- Crank, J. *The Mathematics of Diffusion* (Oxford Univ. Press, New York, 1975).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank F. Widdel and B. Barker Jørgensen for reviewing this manuscript; E. Llobet-Brossa and A. Boetius for discussions; the Hydra Institute for Marine Sciences on Elba for logistical support during collection of the worms; and J. Wulf, D. Lange, G. Eickert and A. Eggers for technical support. The Max-Planck Society and the German Research Foundation (DFG) provided financial support.

Correspondence and requests for materials should be addressed to N.D. (e-mail: ndubilier@mpi-bremen.de). GenBank accession numbers: 16S rRNA: γ -Proteobacteria symbiont AF328856, δ -Proteobacteria symbiont AF328857; DSR: δ -Proteobacteria symbiont AF244995, *D. variabilis* AF191907.

Reproductive isolation caused by colour pattern mimicry

Chris D. Jiggins^{*†}, Russell E. Naisbit^{*}, Rebecca L. Coe[‡] & James Mallet^{*†}

^{*} The Galton Laboratory, University College London, 4 Stephenson Way, London NW1 2HE, UK

[†] Smithsonian Tropical Research Institute, Apartado 2072, Balboa, Panama

[‡] Downing College, The University of Cambridge, Cambridge CB2 1DQ, UK

Speciation is facilitated if ecological adaptation directly causes assortative mating¹, but few natural examples are known. Here we show that a shift in colour pattern mimicry was crucial in the origin of two butterfly species. The sister species *Heliconius melpomene* and *Heliconius cydno* recently diverged to mimic different model taxa, and our experiments show that their mimetic coloration is also important in choosing mates. Assortative mating between the sister species means that hybridization is rare in nature, and the few hybrids that are produced are non-mimetic, poorly adapted intermediates. Thus, the mimetic shift has caused both pre-mating and post-mating isolation. In addition, individuals from a population of *H. melpomene* allopatric to *H. cydno* court and mate with *H. cydno* more readily than those from a sympatric population. This suggests that assortative mating has been enhanced in sympatry.

Mimicry is viewed mainly as a clear, visual demonstration of natural selection within species. But this was not always so: mimicry among Amazonian butterflies was originally presented as a striking example of speciation due to natural selection². More recently, it has been argued that divergence in mimetic pattern can result in intermediates having low fitness because they are non-mimetic and, if colour pattern is also used in mate recognition, assortative mating. Therefore, both pre-mating and post-mating reproductive isolation might result from the evolution of mimicry^{3–5}. Here we study mate choice in *Heliconius* butterflies, a group well known for Müllerian mimicry (mimicry between distasteful species)^{2,4,5}. Closely related *Heliconius* species generally differ in mimetic colour pattern, as though adaptive radiation has occurred^{6,7}. The sister species *H. melpomene* and *H. cydno* are sympatric throughout Central America and the Andean foothills, where they differ in mimicry (Fig. 1) and habitat use⁸. They occasionally hybridize and backcross in nature: hybrid females are sterile, but males are fertile and can be used in the laboratory to introgress genes between the species^{8–10}. In most areas, *H. melpomene* mimics the black, red and yellow pattern of *H. erato*, whilst *H. cydno* mimics the black and white pattern of *H. sapho*. *Heliconius cydno* and *H. melpomene* separated in the last 10⁶ years, much more recently than the non-sister species *H. sapho* and *H. erato* (Fig. 1)¹¹. This and other evidence implies that *H. cydno* and *H. melpomene* have diverged to mimic *H. sapho* and *H. erato*, rather than vice versa¹².

Sympatric Panamanian *H. melpomene* and *H. cydno* did not mate with one another in choice experiments (Tables 1 and 2), although they will do so in no-choice tests^{8–10}. Males from sympatric populations spent over 25 times longer courting virgin females of their own race than heterospecifics (Fig. 2). *Heliconius* females mate soon after eclosion, when they are unable to reject males, so that courtship and assortative mating is largely due to male choice⁷. To test whether males use mimetic colour pattern as a cue in choosing mates, we investigated the response of males to moving models made with either natural wings or coloured paper. Panama *H. melpomene* males approached *H. cydno* colour patterns about half as frequently as those of their own type, and were much less likely (2–4%) to court them (Fig. 3). Similarly, *H. cydno* males were a third as likely to court a *H. melpomene* pattern as their own type, although

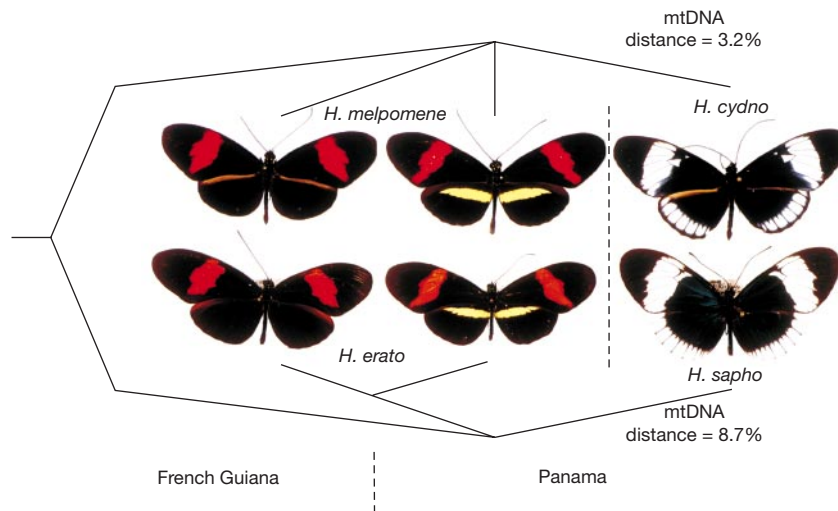


Figure 1 *Heliconius melpomene melpomene* (left, French Guiana), *H. melpomene rosina* (centre, Panama), *H. cydno chioneus* (right, Panama) are shown together with co-mimics (below) *H. erato hydra*, *H. erato cf. petiveranus* and *H. sapho sapho* respectively. Molecular phylogenies (enclosing butterflies) show that the two races of *H. melpomene* and *H. cydno* form an unresolved trichotomy. Mitochondrial sequences suggest *H. melpomene* is paraphyletic with respect to *H. cydno*¹¹, whereas unpublished sequences from nuclear loci show reticulate or mutually monophyletic relationships

between the two species (V. Bull and M. Beltrán, personal communication). Divergence between mitochondrial sequences of *H. erato* and *H. sapho* is almost three times that between *H. melpomene* and *H. cydno* (percentage distance across 940 base pairs of the COI, leu-tRNA and COII genes)¹¹, suggesting that *H. melpomene* and *H. cydno* diverged to mimic *H. erato* and *H. sapho* rather than vice versa. mtDNA, mitochondrial DNA. COI, Cytochrome oxidase I; COII, cytochrome oxidase II.

the probability of initial approach did not differ from that towards conspecifics (Fig. 3). The initial attraction of male *H. cydno* to the red *H. melpomene* pattern may be due to a generalized attraction of *Heliconius* to red flowers. The butterflies clearly responded to visual cues in these experiments, as neither attraction nor courtship differed significantly in comparisons between paper models and real butterfly wings (Fig. 3).

Heliconius melpomene males from French Guiana, where *H. cydno* does not occur, courted live *H. cydno* females twenty times more vigorously than *H. melpomene* males from sympatry with *H. cydno* in Panama (Fig. 2), and the mating experiments showed a similar trend ($G_1 = 3.78$, $P \approx 0.06$; Table 2). This was again a response to colour pattern, as *H. melpomene* males from French Guiana were also more likely than Panamanian *H. melpomene* males to court a *H. cydno* model (combined results from the coloured model and real wing experiments; $G_1 = 8.02$, $P < 0.01$; Fig. 3). In addition, *H. melpomene* males from Panama only reluctantly courted live French Guianan *H. melpomene* females, whereas French Guianan *H. melpomene* males showed no discrimination (Fig. 2); indeed all French Guiana $\times$ Panama *H. melpomene* matings in these tests involved French Guiana males (Table 1). Among *H. melpomene*

racess, males showed greater discrimination between live females (Fig. 2 and Table 1) than between models (Fig. 3), indicating that cues other than colour pattern, such as pheromones, may be involved. Hence, *H. melpomene* males sympatric with *H. cydno* discriminated more strongly than *H. melpomene* allopatric to *H. cydno*. This pattern is expected if mate preference has been 'reinforced' to prevent the production of unfit hybrid offspring in sympatry^{13,14}, although the evidence would be strengthened if replicated with other allopatric and sympatric populations^{14,15}. Of course, character displacement between non-hybridizing species cannot be ruled out. For example, the presence of *H. sapho* might

Table 1 Number of matings in tetrad mate choice experiments

Female	Male	Male
Sympatric populations		
	<i>H. melpomene</i> (Panama)	<i>H. cydno</i> (Panama)
<i>H. melpomene</i> (Panama)	14	0
<i>H. cydno</i> (Panama)	0	11
Allopatric populations		
	<i>H. melpomene</i> (Panama)	<i>H. melpomene</i> (Guiana)
<i>H. melpomene</i> (Panama)	9.5	4
<i>H. melpomene</i> (Guiana)	0	13.5
	<i>H. melpomene</i> (Guiana)	<i>H. cydno</i> (Panama)
<i>H. melpomene</i> (Guiana)	14.5	0
<i>H. cydno</i> (Panama)	3	12.5

Mating results of 0.5 are due to two cases of virtually simultaneous mating by both pairs in a tetrad (see Methods).

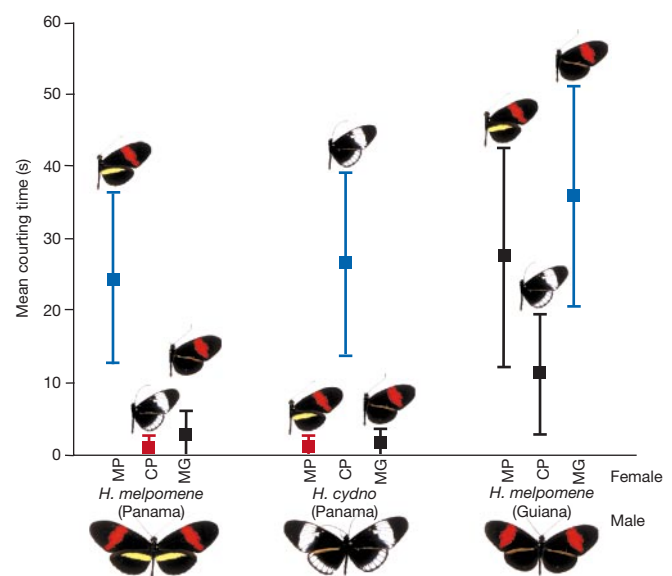


Figure 2 Time spent courting live females in 10-min trials with 95% confidence intervals. Red, comparisons between sympatric populations; black, comparisons between allopatric populations; blue, comparisons between males and females of the same genotype. MP, *H. melpomene* (Panama); CP, *H. cydno* (Panama); MG, *H. melpomene* (Guiana).

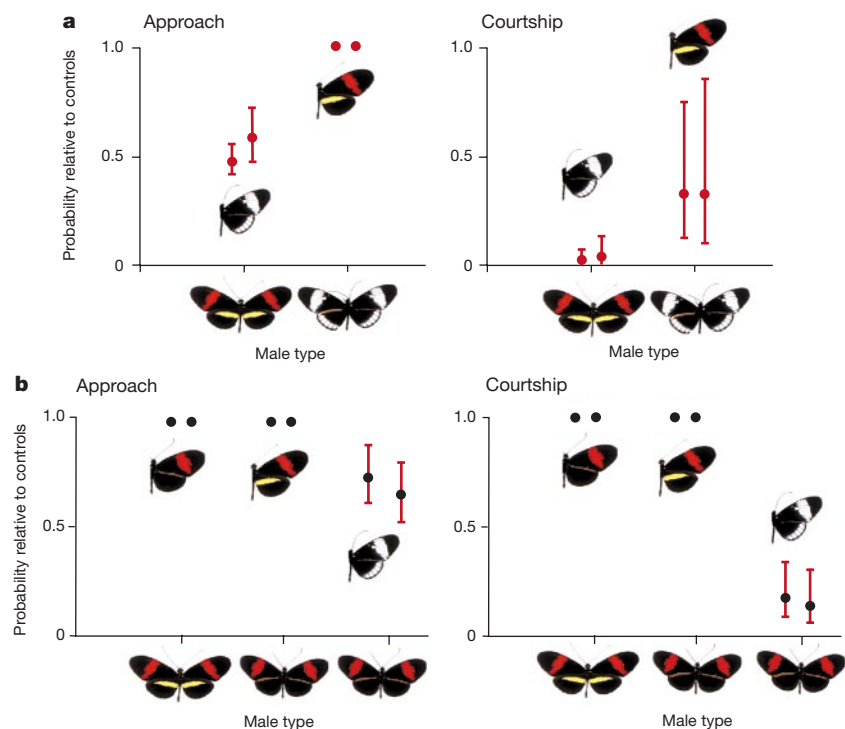


Figure 3 Relative probabilities of male approach and courtship of colour pattern models. Comparisons between Panama populations (sympatry) (**a**) and with the Guiana population (allopatry) (**b**). Values are estimated relative to within-race controls (equal to 1 in each case). Paired data points for experiments using real wings (left) and coloured paper

models (right) are shown for each comparison. Values of Q_A (approach) and Q_H (hovering courtship) were estimated with support limits under the ten-parameter model. Setting real wing and paper model parameters equal gives no significant reduction of fit ($G_5 = 3.70$).

also lead to enhanced rejection of the shared *Heliconius sapho*/*H. cydno* pattern by *H. melpomene*. Here we show that mimetic colour patterns are also important in mate recognition. Assortative mating contributes to speciation because post-mating isolation between *H. melpomene* and *H. cydno* is incomplete^{8,9}. As in inter-racial hybrid zones, intermediate colour patterns are unlikely to be recognized as distasteful by predators, generating strong disruptive selection¹⁶. Selection on mimicry may be strong, $S \approx 0.2-0.3$ per locus in inter-racial hybrid zones giving $S \approx 0.6$ overall^{8,16}, comparable to that caused by F_1 female sterility ($S \approx 0.5$). Mimicry therefore provides an example of a trait under strong ecological selection that is also used as a mating cue. Such pleiotropy between mate choice and disruptive selection is an important feature of speciation theory, because it can trigger rapid speciation with a high probability^{1,17,18}, but only a few other examples are known¹⁹⁻²¹. The great diversity of colour pattern races within many *Heliconius* species shows that mimetic shifts rarely lead to speciation. Only where a shift dramatically changes colour or appearance, such as that between *H. melpomene* and *H. cydno*, will mate choice co-evolve sufficiently with mimicry to generate pre-mating isolation.

In addition, *H. melpomene* and its co-mimic *H. erato* occur in light gaps and secondary forest, whereas *H. cydno* and *H. sapho* are found in more primary forest, albeit with considerable overlap^{8,22}. This habitat shift, associated with mimicry, will itself contribute further to pre-mating isolation. In conclusion, pre- and post-mating isolation between *H. melpomene* and *H. cydno* has resulted from an adaptive shift in ecology and mimicry, in association with partial hybrid sterility. Subsequently, assortative mating between sympatric populations has become enhanced, possibly owing to reinforcement. This and other recent examples suggest that ecological adaptation can result in assortative mating as a byproduct and may be an important and largely overlooked cause of speciation¹⁹⁻²¹. □

Methods

Heliconius melpomene melpomene were collected near Cayenne, French Guiana in May 1998 (around 35 individuals) and February 1999 (58 individuals). *Heliconius cydno chioneus* and *H. melpomene rosina* were obtained continually from near Gamboa, Panama. Experiments were performed with descendants (three or fewer generations after collection), in insectaries⁷ in Gamboa in 1998–1999.

Mate choice experiments

‘Tetrad’ experiments, consisting of a recently emerged virgin female (1 day old or less) and a mature male (more than 5 days old) of each of two genotypes, were performed in $1 \times 1 \times 2$ m insectaries. The first mating was recorded for each experiment; individuals were not reused. On two occasions, both pairs mated simultaneously and so were scored as each having 0.5 matings. At least 25 experiments were performed per comparison. Likelihood was used to estimate the probability $P_{i \times j}$ of a mating⁷ between female type i and male type j , relative to $P_{mp \times mg}$ of a mating within Guiana *H. melpomene* (MG), which was set to 1. The overall multinomial probability of the results for each experiment were then estimated, e.g. for the Panama *H. melpomene* (MP) $\times$ Panama *H. cydno* (CP) comparison, $\Phi_{mp \times mp} = P_{mp \times mp} / (P_{mp \times mp} + P_{mp \times cp} + P_{cp \times mp} + P_{cp \times cp})$, $\Phi_{mp \times cp} = P_{mp \times cp} / (P_{mp \times mp} + P_{mp \times cp} + P_{cp \times mp} + P_{cp \times cp})$ and so on. ($\Sigma \Phi = 1$ for each tetrad). The log likelihood was therefore $\Sigma (X_{mp \times mp} \log \Phi_{mp \times mp} + X_{mp \times cp} \log \Phi_{mp \times cp} + \dots)$ where $X_{mp \times mp}$ is the number of MP $\times$ MP matings and $X_{mp \times cp}$ the number of MP $\times$ CP matings in that tetrad. Likelihoods were summed over all experiments and maximized by

Table 2 Relative probabilities of mating between genotypes			
Female	Male	Male	Male
	<i>H. melpomene</i> (Panama)	<i>H. cydno</i> (Panama)	<i>H. melpomene</i> (Guiana)
<i>H. melpomene</i> (Panama)	1	0 (0, 0.167)	0.348 (0.099, 0.921)
<i>H. cydno</i> (Panama)	0 (0, 0.167)	1	0.222 (0.051, 0.641)
<i>H. melpomene</i> (Guiana)	0 (0, 0.182)	0 (0, 0.154)	[1]

Probabilities were estimated relative to *H. melpomene* (Guiana) $\times$ *H. melpomene* (Guiana), which was set to 1 (square brackets). Support limits are shown in parentheses (values of 1 without support limits are not significantly different from 1).

varying P_{ixj} values. Setting $P_{ixj} = 1$ (within all genotypes) did not significantly reduce the fit ($G_2 = 0.81$, not significant), suggesting similar mating propensity among genotypes. Asymmetries ($P_{ixj} \neq P_{jxi}$) can therefore be presumed to be due to mate choice rather than mating propensity. Parameters and support limits (asymptotically equivalent to 95% confidence intervals²³) were estimated under the simpler six-parameter model.

Live female courtship experiments

We placed two or three males (more than 5 days old) of different genotypes in an insectary and introduced a single virgin female (1–5 days old). Courtship (sustained hovering by the male over the female) was recorded over a period of 10 min. The female genotype was then substituted, with genotype order randomized. On mating, pairs were quickly and gently separated, which did not disrupt subsequent behaviour. Males were never reused, but females were drawn randomly from a pool of three to four individuals per genotype. In all, 840 min of observations were made in 19 replicates with all three male genotypes and a further nine with MP and MG males alone.

Colour pattern models

Between five and fifteen males in a $2 \times 2 \times 2$ m insectary were presented with dissected natural wings or a colour pattern model, fixed to a length of flexible wire on a lightweight handle. Models were manipulated to simulate *Heliconius* flight in the centre of a spherical area (60 cm diameter) demarcated by a bamboo cross. Randomly ordered pairs of 5-min experiments were carried out: (1) a control flight with a model of the male's own colour pattern and (2) an experimental flight with a different colour pattern. Entry to the sphere was recorded as 'approach' and sustained fluttering directed at the model as 'courtship'. At least ten replicates were carried out per comparison. Each procedure was repeated with real female wings and paper models colour-matched using commercially available permanent marker pens. Reflectance spectra of real and paper models were similar (Supplementary Information), and male behaviour towards wings and models did not differ significantly (see below).

Numbers of approaches (X_A) and hovering courtship interactions (X_H) are given in the Supplementary Information. We estimated the probabilities Q_{ixj} that males of type j approached or courted models of type i relative to that of their own type j , using likelihood. Thus, for MP males with MP versus CP models, the actual probabilities are $Q_{Acp \times mp} / (Q_{Acp \times mp} + 1)$ that males approach CP and $1 / (Q_{Acp \times mp} + 1)$ that they approach MP. The log_e likelihood for this experiment is therefore $\sum [X_{Acp \times mp} \log_e \{Q_{Acp \times mp} / (Q_{Acp \times mp} + 1)\} + X_{Hmp \times mp} \log_e \{1 / (Q_{Acp \times mp} + 1)\}]$, where $X_{Acp \times mp}$ is the number of MP males approaching CP and $X_{Hmp \times mp}$ is the number approaching MP. Similarly Q_{Hixj} parameters were estimated for probability of hovering courtship of the model. Estimates were obtained for paper models as well as real wings, giving a total of 20 parameters. The summed log_e likelihood was maximized over all experiments by varying the Q_{ixj} parameters. Subsequently, all comparisons within *H. melpomene* and $Q_{Amp \times cp}$ parameters were set to 1 without loss of fit ($G_{10} = 11.02$, not significant). Parameter values for the resultant ten-parameter model are shown in Fig. 3. Real and paper model parameters do not differ significantly ($G_5 = 3.70$), giving a combined five-parameter model.

Received 14 July 2000; accepted 1 February 2001.

- Maynard Smith, J. Sympatric speciation. *Am. Nat.* **100**, 637–650 (1966).
- Bates, H. W. Contributions to an insect fauna of the Amazon valley. Lepidoptera: Heliconidae. *Trans. Linn. Soc. Lond.* **23**, 495–566 (1862).
- Vane-Wright, R. I. in *Diversity of Insect Faunas* (eds Mound, L. A. & Waloff, N.) 56–70 (Blackwell Scientific, Oxford, 1978).
- Turner, J. R. G. Adaptation and evolution in *Heliconius*: a defense of neo-Darwinism. *Annu. Rev. Ecol. Syst.* **12**, 99–121 (1981).
- Mallet, J. & Joron, M. Evolution of diversity in warning colour and mimicry: Polymorphisms, shifting balance and speciation. *Annu. Rev. Ecol. Syst.* **30**, 201–233 (1999).
- Turner, J. R. G. Adaptive radiation and convergence in subdivisions of the butterfly genus *Heliconius* (Lepidoptera: Nymphalidae). *Zool. J. Linn. Soc.* **58**, 297–308 (1976).
- McMillan, W. O., Jiggins, C. D. & Mallet, J. What initiates speciation in passion vine butterflies? *Proc. Natl Acad. Sci. USA* **94**, 8628–8633 (1997).
- Mallet, J., McMillan, W. O. & Jiggins, C. D. in *Endless Forms: Species and Speciation* (eds Howard, D. J. & Berlocher, S. H.) 390–403 (Oxford Univ. Press, New York, 1998).
- Linares, M. *Adaptive Microevolution Through Hybridization and Biotic Destruction in the Neotropics* (Thesis, Univ. Texas, Austin, Texas, 1989).
- Gilbert, L. E. in *Ecology and Evolution Taking Flight: Butterflies as Model Systems* (eds Boggs, C. L., Watt, W. B. & Ehrlich, P. R.) (Univ. Chicago Press, Chicago, Illinois, in the press).
- Brower, A. V. Z. Parallel race formation and the evolution of mimicry in *Heliconius* butterflies: a phylogenetic hypothesis from mitochondrial DNA sequences. *Evolution* **50**, 195–221 (1996).
- Mallet, J. Causes and consequences of a lack of coevolution in Müllerian mimicry. *Evol. Ecol.* **13**, 777–806 (1999).
- Dobzhansky, T. Speciation as a stage in evolutionary divergence. *Am. Nat.* **74**, 312–321 (1940).
- Butlin, R. Speciation by reinforcement. *Trends Ecol. Evol.* **2**, 8–12 (1987).
- Noor, M. A. F. Reinforcement and other consequences of sympatry. *Heredity* **83**, 503–508 (1999).
- Mallet, J. & Barton, N. H. Strong natural selection in a warning color hybrid zone. *Evolution* **43**, 421–431 (1989).
- Dieckmann, U. & Doebeli, M. On the origin of species by sympatric speciation. *Nature* **400**, 354–357 (1999).
- Kirkpatrick, M. Reinforcement and divergence under assortative mating. *Proc. Roy. Soc. Lond. B* **267**, 1649–1655 (2000).
- Craig, T. P., Itami, J. K., Abrahamson, W. G. & Horner, J. D. Behavioral evidence for host-race formation in *Eurosta solidaginis*. *Evolution* **47**, 1696–1710 (1993).
- Feder, J. L. et al. Host fidelity is an effective premating barrier between sympatric races of the apple

maggot fly. *Proc. Natl Acad. Sci. USA* **91**, 7990–7994 (1994).

- Rundle, H. D., Nagel, L., Wenrick Boughman, J. & Schluter, D. Natural selection and parallel speciation in sympatric sticklebacks. *Science* **287**, 306–308 (2000).
- Mallet, J. & Gilbert, L. E. Why are there so many mimicry rings? Correlations between habitat, behaviour and mimicry in *Heliconius* butterflies. *Biol. J. Linn. Soc.* **55**, 159–180 (1995).
- Edwards, A. W. F. *Likelihood* (Cambridge Univ. Press, Cambridge, 1972).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank C. Paresce, C. Estrada, R. Woods, M. Beltrán and L. O'Donovan for help with experiments; B. Hermiers for help in French Guiana; F. Jiggins, M. Joron, I. Emelianov, E. Bermingham and G. Hurst for discussion; and S. Upson at Perkin Elmer for reflectance measurements. This work was carried out with permission from ANAM (Autoridad Nacional del Ambiente) to work in the República de Panamá, and was funded by the Natural Environment Research Council and a BBSRC (Biotechnology and Biological Sciences Research Council) studentship.

Correspondence and requests for materials should be addressed to C.D.J. (e-mail: jigginsc@naos.si.edu).

Lesions of the human amygdala impair enhanced perception of emotionally salient events

Adam K. Anderson^{†*} & Elizabeth A. Phelps[‡]

^{*} Department of Psychology, Yale University, New Haven, Connecticut 06520, USA

[‡] Department of Psychology, New York University, New York, New York 10003, USA

Commensurate with the importance of rapidly and efficiently evaluating motivationally significant stimuli, humans are probably endowed with distinct faculties^{1,2} and maintain specialized neural structures to enhance their detection. Here we consider that a critical function of the human amygdala^{3,4} is to enhance the perception of stimuli that have emotional significance. Under conditions of limited attention for normal perceptual awareness—that is, the attentional blink^{5,6}—we show that healthy observers demonstrate robust benefits for the perception of verbal stimuli of aversive content compared with stimuli of neutral content. In contrast, a patient with bilateral amygdala damage has no enhanced perception for such aversive stimulus events. Examination of patients with either left or right amygdala resections shows that the enhanced perception of aversive words depends specifically on the left amygdala. All patients comprehend normally the affective meaning of the stimulus events, despite the lack of evidence for enhanced perceptual encoding of these events in patients with left amygdala lesions. Our results reveal a neural substrate for affective influences on perception, indicating that similar neural mechanisms may underlie the affective modulation of both recollective^{7–9} and perceptual experience.

The amygdala supports substantial projections to primary and higher-order sensory areas and the hippocampal formation¹⁰. Thus, the amygdala is strategically placed to allow emotional value^{4,11} both to modulate perceptual sensitivity to incoming information and to bolster its post-encoding consolidation into memory. Much evidence has shown that the amygdala is involved in the latter of these

[†] Present address: Department of Psychology, Stanford University, Jordan Hall, Building 420, Stanford, California 94305-2130, USA

two functions, as it is critical for enhanced memory of emotionally arousing events^{7–9,12–14}. It is unknown, however, whether the human amygdala also provides a means for affective value to modulate perception itself, thus influencing the likelihood that stimulus events reach awareness.

We examined the importance of the amygdala in the affective modulation of perception by using the attentional-blink effect^{5,6}. The attentional blink shows that, after identifying a single target stimulus, there is a transient impairment in awareness for a subsequently presented second target^{5,6,15}. Consistent with the notion of a differential sensitivity toward significant events, this deficit in perceptual awareness is greatly attenuated for aversive verbal stimuli¹⁶. Considering that the comprehension of meaning from word forms is left largely intact after amygdala lesions^{8,17,18}, the use of linguistic events allows analysis of the influence of affective content on perceptual encoding, independent of a primary deficit in appreciation of emotional significance. Here we examined the attenuation of the attentional blink for affectively significant words in a patient with bilateral damage to the amygdala (S.P.)⁹, and in 10 patients with unilateral lesions of the right or left amygdala.

Observers were asked to report the identity of two green target words occurring amongst a stream of black distractor items during rapid serial visual presentation (RSVP) (Fig. 1). Confirming the affective modulation of the attentional blink, control observers identified negative words with greater accuracy than neutral words across all temporal lags (79.8% versus 61.5%; $P < 0.0001$). A significant interaction between T1–T2 temporal lag (early versus late) and T2 valence (negative versus neutral) ($P < 0.02$) indicated that the degree of affective modulation of the attentional blink was most pronounced at early lags ($P < 0.0001$), where attentional resources were most occupied with the processing of the preceding T1.

Similar to controls, S.P. showed a normal attentional blink as evidenced by her identification accuracy increasing from early relative to later temporal lags ($P < 0.0001$). S.P.'s performance also recovered to baseline levels over a normal time course; she exhibited similar levels of T2 identification performance to controls (controls 80.4% versus S.P. 81.0%) on the longest T1–T2 lag (stimulus onset asynchrony; SOA = 910 ms). Unlike control observers, S.P. showed almost no advantage for negative compared with neutral word identification during the middle of the attentional blink at early T1–T2 temporal lags (25.0% versus 23.0%; $P > 0.80$; Fig. 2b); she fell significantly below the control range in identifying negative items at early lags (controls 72.6%, versus S.P. 25.0%; $z = 2.11$, $P < 0.02$). This impairment in identifying negative events was most evident when T2 was presented immediately after T1 (controls 74.5%, versus S.P. 12.5%; $z = 3.13$, $P < 0.001$). Furthermore, an analysis of the distribution of difference scores (negative minus neutral) showed that whereas the advantage for negative words was most prominent in controls at early lags, S.P. showed almost no advantage for negative words falling outside the control range (controls 22.6%, versus S.P. 2.1%; $z = 1.76$, $P < 0.04$; Fig. 2d).

To control for any effect of lower baseline performance in S.P., we examined one half of the control subjects ($n = 10$) that performed similarly to S.P. on neutral items (controls 47.3% versus S.P. 43.8%; $z = 0.13$). In contrast to S.P., these controls had a highly robust affective modulation of the attentional blink (negative 70.6% versus neutral 47.3%; $P < 0.0001$; Fig. 2b). An examination of the advantage of negative-item identification (negative minus neutral) again showed an impairment in affective modulation at early temporal lags (controls 26.2% versus S.P. 2.1%; $z = 1.72$; $P < 0.05$), with controls showing a more substantial proportional increase in identifying negative events compared with S.P. (83.2% versus 8.7%).

The observed impaired affective modulation of the attentional blink at early temporal lags may reflect S.P.'s difficulty in maintaining items in her memory over the interval before report. Arguing

against this notion, however, her performance in identifying T1, which occurred before T2, was even more accurate than that of controls at the longest T1–T2 lag (96.8% versus 91.5%).

An additional concern was that S.P.'s deficit was not due to impaired affective modulation of perception, but might reflect a more global deficit that disrupts any type of modulation of perceptual encoding. However, when we manipulated the perceptual discriminability of neutral targets by altering the visual similarity of the targets and distractors⁶, S.P.'s word identification performance was greatly enhanced (high discriminability 78.1%, versus low discriminability 52.1%; $P < 0.001$; Fig. 2c). The influence of perceptual salience was most pronounced at early T1–T2 lags (high discriminability 66.7%, versus low discriminability 31.2%; a proportional identification benefit of over 100%). A similar analysis of S.P.'s performance in the affective salience condition showed no reliable overall advantage (negative 50.8% versus neutral 44.6%; $P > 0.41$), or a lag-dependent advantage for negative words ($P > 0.49$). This finding indicates that affective and perceptual modulatory influences on visual awareness may be dissociable, with only the latter being dependent on the amygdala.

Functional neuroimaging studies of word perception find that the presentation of emotional linguistic stimuli may specifically activate the amygdala in the left hemisphere¹⁹, suggesting that S.P.'s selective left amygdala damage in particular may be responsible for the observed deficit. This would be consistent with evidence that verbal/nonverbal hemispheric asymmetries typically ascribed to higher cortical function in humans extend to the function of subcortical structures such as the amygdala^{20–22}. To replicate the

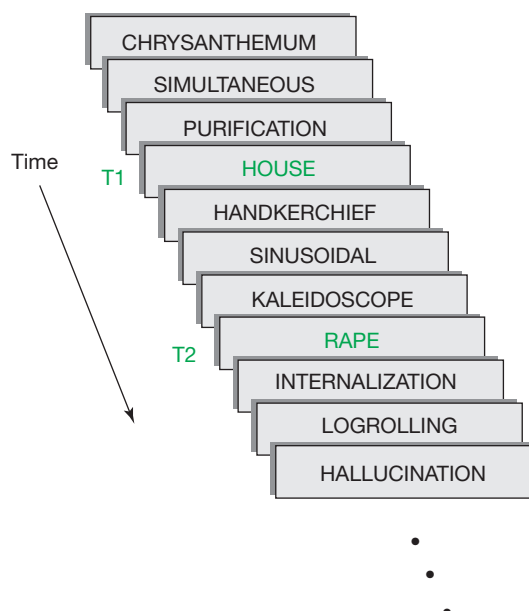


Figure 1 Diagram of the dual-target rapid serial visual presentation (RSVP) task. Fifteen words were briefly presented sequentially in an identical central location, and observers were instructed to ignore words appearing in black (distractors) and indicate the identity of two target words occurring in green. Distractor words were longer in length to ensure sufficient masking of the target events. Responses were made by typing on a computer keyboard immediately after the stimulus sequence. The temporal lag between the first (T1) and second target (T2) was varied. T2 items occurring during early T1–T2 temporal lags (< 4 intervening distractor items, < 600 ms) are susceptible to the attentional blink (attentional blink), an attention-related impairment in perceptual encoding. As attentional resources become more available during late T1–T2 temporal lags (> 4 intervening distractor items, > 600 ms), T2 items are less susceptible to the attentional blink. In the described experiments, the affective value of T2 was manipulated and T2 identification accuracy was the critical measure.

finding of impaired affective modulation of perceptual encoding in a larger patient sample, and to assess more precisely which particular neural structures damaged in S.P. were critical for this deficit, we examined five patients with right (RTL) and five with left (LTL) amygdala lesions due to unilateral anteromedial temporal lobe resections.

When RTL and LTL observers were asked to identify neutral and negative T2 words during dual-target RSVP, we found that the effect of stimulus valence on identification was not equivalent across

patient and control groups ($P < 0.02$). Similar to controls ($P > 0.27$), RTL patients showed a highly significant affective modulation of the attentional blink (negative 68.9% versus neutral 54.6%; $P < 0.009$), which was most evident at early T1–T2 lags (negative 65.7% versus neutral 44.4%; Fig. 3a). In contrast, LTL patients showed no significant advantage in negative-event identification (negative 41.1% versus neutral 36.2%; $P > 0.16$), demonstrating impaired affective modulation of the attentional blink relative to controls ($P < 0.005$).

When we examined the extent of the negative-event advantage

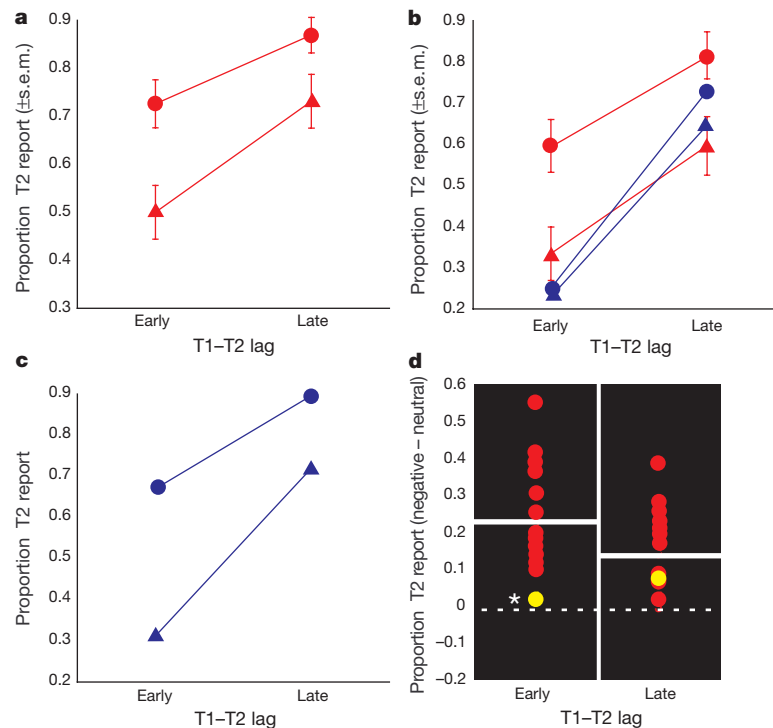


Figure 2 Proportion of T2 items correctly identified at early and late T1–T2 temporal lags. **a**, Controls ($n = 20$; red) showed enhanced identification of negative (circles) compared with neutral (triangles) items during the attentional blink. **b**, S.P. (blue) showed no evidence of enhanced identification of negative (circles) compared with neutral (triangles) items during the attentional blink, and remained significantly impaired when compared with the half of the controls ($n = 10$, in red) that showed similar baseline performance on neutral items. Bars in **a** and **b** show s.e.m. **c**, Impaired affective modulation of perception in S.P. did not generalize to the influence of non-affective manipulations of stimulus

salience on perceptual awareness. S.P. exhibits a large benefit for high (circles) compared with low (triangles) salience T2 events. **d**, Distribution of T2 difference scores (negative minus neutral) for controls ($n = 20$) and S.P. at early and late T1–T2 temporal lags. Yellow, S.P.; red, controls; white bar, control mean. In contrast to controls, S.P. did not show a reliable advantage for identifying negative compared with neutral items at early T1–T2 temporal lags, falling significantly outside the control range (controls 22.6%, S.P. 2.1%). Asterisk denotes significant impairment.

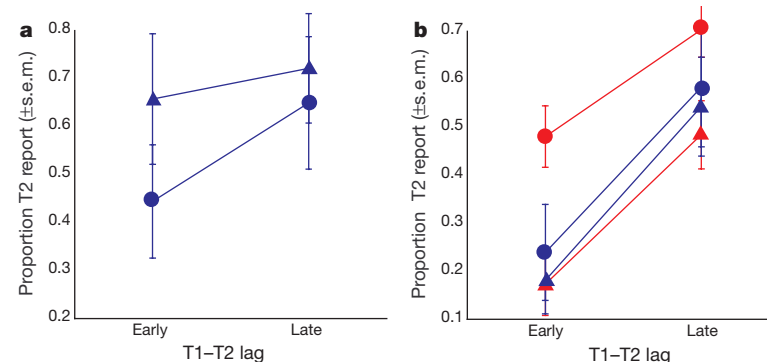


Figure 3 Proportion of T2 correctly identified for negative (circles) and neutral (triangles) items at early and late T1–T2 temporal lags for RTL and LTL patients. **a**, RTL patients showed enhanced identification, to a similar extent as controls, of negative compared with neutral words (see Fig. 2a). **b**, LTL patients (blue) showed no evidence of enhanced

identification of negative compared with neutral items, and remained significantly impaired when compared with controls (red) that showed similar baseline performance on neutral items. Bars in **a** and **b** show s.e.m.

(negative minus neutral), we found that it depended on subject group ($P < 0.03$). Both controls ($P < 0.007$) and the RTL group ($P < 0.05$) showed a significant valence-dependent advantage in stimulus identification compared with the LTL group (controls 22.6%, RTL 21.3%, LTL 5.8%). There was no significant difference between the RTL and control groups ($P > 0.80$). When we precisely matched LTL patients ($n = 5$) with control observers showing equivalent baseline responding on neutral items ($n = 5$) (LTL 36.2%, controls 32.7%), LTL patients remained differentially impaired on negative-event identification ($P < 0.003$; Fig. 3c). In contrast, an analysis of performance on the preceding T1 showed that there were no significant differences between the three groups, and near numerically identical performance between RTL and LTL patients (controls 89.6%, RTL 79.6%, LTL 80.9%; $P > 0.27$).

The impaired perceptual encoding advantage for stimuli of affective significance may be only secondary to impaired comprehension of the affective importance of these events. To investigate this hypothesis, we considered the evaluation of each T2 item used in the affective salience RSVP task according to two stimulus dimensions: valence and arousal. Control participants rated negative words as both more negative (4.61 ± 0.86 versus 3.28 ± 0.73 ; $P < 0.0001$) and more arousing (4.16 ± 1.5 versus 2.06 ± 1.03 ; $P < 0.0001$) than the neutral words. S.P.'s ratings similarly discriminated between the valence (negative 4.79 versus neutral 4.04) and arousal value (negative 4.82 versus neutral 1.75) of negative and neutral words, falling in the normal control range for each ($P > 0.15$ in all cases). Likewise, RTL and LTL patients exhibited intact discrimination between the valence of the target words (RTL_{negative} 5.04 versus RTL_{neutral} 2.78; LTL_{negative} 5.17 versus LTL_{neutral} 2.38), with both patient groups showing an even greater polarization between the appreciation of items of negative and neutral valence ($P < 0.04$). In addition, RTL and LTL patients provided normal ratings of arousal value (RTL_{negative} 2.96 versus RTL_{neutral} 1.34; LTL_{negative} 3.75 versus LTL_{neutral} 2.44; $P > 0.40$). Thus, the impaired influence of affective content on perceptual awareness shown in S.P. and LTL patients was not a reflection of an impaired comprehension of the affective attributes of these same stimuli. This result is similar to evidence that patients with amygdala lesions can exhibit normal subjective impressions of emotional stimuli, but show little enhanced memory for these events⁷⁻⁹.

Related to our study, patients with amygdala lesions exhibit normal subjective and psychophysiological responses to aversive word types, but do not show the normal retention advantage associated with these words^{8,9}. Consistent with animal models of the amygdala's role in supporting perceptual vigilance^{23,24}, our findings indicate that neuromodulatory influences of the human amygdala may not be restricted to episodic memory consolidation, but may extend to initial perceptual encoding^{25,26}, influencing the likelihood that events of affective importance reach awareness.

Supporting the notion that the human amygdala is involved in the perceptual encoding of emotionally significant linguistic events, visually presented negative-word stimuli evoke greater activity in the amygdala compared with words of neutral value^{19,27}. Although such results may suggest that the amygdala itself directly encodes linguistic events, we propose that lesions of the amygdala disrupt its ability to modulate the efficiency of word processing that takes place in other brain regions, such as the posterior fusiform gyrus^{19,27}. Consistent with its larger perceptual neuromodulatory role, anatomical studies of primate amygdala projections show widespread influences throughout all levels of the ventral visual processing pathway¹⁰, allowing the amygdala to influence cortical modules that are specialized for the processing of different informational domains.

The exact neural mechanisms by which the amygdala influences the salience of stimulus events processed in target perceptual systems remain unclear. The amygdala may modulate the fluency of perceptual encoding either by influencing processing dynamics in

perceptual systems by transiently modulating cortical firing thresholds^{23,25}, or by influencing neuronal plasticity by altering the receptive field properties of subcortical and primary and secondary cortical regions^{26,28}. Our results suggest that the result of these amygdala-mediated neural modifications may be to enhance perceptual sensitivity to events of importance to the organism, making them less dependent on attentional resources to achieve awareness. Thus, one of the critical functions of the human amygdala is to segregate the neural representations of the significant from the mundane²⁹, either through later recollective processes^{7-9,12-14} or during stimulus encoding, shaping perceptual experience directly. □

Methods

Participants

S.P. is a 54-year-old right-handed female who, at the age of 48, had her right amygdala removed as a result of anteromedial temporal lobe resection for medically intractable epilepsy. Her right temporal lobe resection included partial removal of the anterior, middle and inferior temporal gyri, and complete removal of the hippocampus, parahippocampus and projection fibres to their posterior extent. Before surgery, an additional lesion was observed in the left amygdala region. Two biopsies in this region revealed reactive gliosis consistent with mesial temporal sclerosis. Post-surgery T1- and T2-weighted magnetic resonance scans (see Supplementary Information) show abnormal signal intensity extending throughout her left amygdala. Neuropsychological and radiological indices suggest that the damage does not extend to adjacent temporal lobe structures in the left hemisphere⁹. S.P. received a high-school education, has taken college courses, and presents a normal neuropsychological profile⁹.

In addition to S.P.'s performance, ten patients with unilateral anteromedial temporal lobectomy (TLB) including removal of the amygdala were examined: five right (RTL) and five left (LTL). Each of these patients had undergone a highly uniform surgical procedure³⁰ (described above) to control for medically intractable seizures. The performance of S.P. and TLB patients (RTL: age, 42.6 ± 7.4 yr; education, 14.4 ± 2.6 yr; LTL: age, 33.4 yr; education, 14.6 ± 2.5 yr) was compared with that of 20 (15 female and 5 male) control participants of similar age (43 ± 7.7 yr) and education (14.1 yr).

Design and procedure

The first target stimuli (T1) were 56 neutral words. In the affective salience RSVP task, the second target stimuli (T2) consisted of 28 negative (for example, rape, bastard) and 28 neutral words (for example, broom, distance). The negative word list comprised aversive words intended to be more negative and physiologically arousing in nature than their neutral counterparts⁸. The negative and neutral lists were matched for average word length, written word frequency and interletter frequency. Distractor words were neutral in content and longer (mean 11.66 letters) than the target stimuli to ensure sufficient masking.

Each trial consisted of 15 word items, 2 targets and 13 distractors. In the perceptual salience control condition, an alphanumeric character mask replaced distractor words and both T1 and T2 were neutral words. T1 and T2 were designated as targets by appearing in bright green, whereas the distractor words appeared in black. Each item in the stream was presented for 130 ms and immediately followed the subsequent item, resulting in seven lags between T1 and T2 ranging from lag 1 (no intervening items, SOA = 130 ms) to lag 7 (SOA = 910 ms). S.P. was tested at the same SOA as controls; TLB patients were tested at a slightly longer SOA of 150 ms. There were 16 trials for each factor combination of lag (1-7) and condition (negative versus neutral), for a total of 224 trials. The subjects' task was to monitor the RSVP stream and then report the two green coloured targets by typing them after the stimulus sequence (Fig. 1). After the dual-target RSVP task, we presented subjects with the T2 words and asked them to evaluate each word on two seven-point scales: valence (1, pleasant; 7, unpleasant) and arousal (1, low; 7, high).

Apparatus

The experiment was conducted on a Macintosh computer and stimuli were presented in Geneva font, point-size 24, viewed from an average distance of 40 cm. The background was a uniform grey field.

Analysis

For the control participant sample, data for the first target (T1) and second target (T2) contingent on correct identification of T1 were considered separately and submitted to two independent 2×2 factorial analyses of variance (ANOVAs), with condition (negative versus neutral) and T1-T2 lag (early (collapsing across lags 1-3) versus late (collapsing across lags 5-7)) entered as independent within-subject factors. Collapsing across independent lags was undertaken to highlight the essential contrast between when attentional resources for perceptual encoding were more available (late lags) versus less available (early lags).

To assess deviation from normal performance, S.P.'s data were converted to z scores on the basis of the control means and standard deviations. To avoid the loss of data points, S.P.'s performance on T2 was calculated independently of T1 performance. S.P.'s data from the affective salience and perceptual salience conditions were also submitted to separate single-subject ANOVAs, in which each trial was entered as the random factor. TLB data for

T1 and T2 contingent on correct identification of T1 were considered separately and submitted to two independent mixed-design ANOVAs. In addition, control, LTL and RTL difference scores (negative minus neutral) were submitted to an additional ANOVA. The valence and arousal evaluation ratings for the control, RTL and LTL groups were submitted to separate between-group *t*-tests.

Received 19 January; accepted 16 March 2001.

1. Bruner, J. S. & Postman, L. Emotional selectivity in perception and reaction. *J. Pers.* **16**, 69–77 (1947).
2. Niedenthal, P. M. & Kitayama, S. (eds) *The Heart's Eye: Emotional Influences in Perception and Attention* (Academic, San Diego, 1994).
3. Aggleton, J. P. *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction* (Wiley-Liss, New York, 1992).
4. LeDoux, J. E. *The Emotional Brain* (Simon & Schuster, New York, 1992).
5. Raymond, J. E., Shapiro, K. L. & Arnell, K. M. Temporary suppression of visual processing in an RSVP task: An attentional blink? *J. Exp. Psychol. Hum. Percept. Perform.* **18**, 849–860 (1992).
6. Chun, M. M. & Potter, M. C. A two-stage model for multiple target detection in rapid serial visual presentation. *J. Exp. Psychol. Hum. Percept. Perform.* **21**, 109–127 (1995).
7. Cahill, L., Babinsky, R., Markowitsch, H. J. & McGaugh, J. L. The amygdala and emotional memory. *Nature* **377**, 295–296 (1995).
8. LaBar, K. S. & Phelps, E. A. Arousal mediated memory consolidation: Role of the medial temporal lobe in humans. *Psychol. Sci.* **9**, 490–493 (1998).
9. Phelps, E. A. *et al.* Specifying the contributions of the human amygdala to emotional memory: a case study. *Neurocase* **4**, 527–540 (1998).
10. Amaral, D. G., Price, J. L., Pitkanen, A. & Charnichael, S. T. In *The Amygdala* (ed. Aggleton, J. P.) 1–66 (Wiley-Liss, New York, 1992).
11. Nishijo, H., Ono, T. & Nishino, H. Single neuron responses in amygdala of alert monkeys during complex sensory stimulation with affective significance. *J. Neurosci.* **8**, 3570–3583 (1988).
12. Cahill, L. *et al.* Amygdala activity at encoding correlated with long-term, free recall of emotional information. *Proc. Natl Acad. Sci. USA* **93**, 8016–8021 (1996).
13. Hamann, S. B., Ely, T. D., Grafton, S. T. & Kilts, C. D. Amygdala activity related to enhanced memory for pleasant and aversive stimuli. *Nature Neurosci.* **2**, 289–294 (1999).
14. Canli, T., Zuo, Z., Brewer, J., Gabrieli, J. D. & Cahill, L. Event-related activation in the human amygdala associates with later memory for individual emotional experience. *J. Neurosci.* **20**, RC99, 1–5 (2000).
15. Vogel, E. K., Luck, S. J. & Shapiro, K. L. Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* **24**, 1656–1674 (1998).
16. Anderson, A. K. Affective influences on attentional dynamics during perceptual encoding: Investigations with the attentional blink. *J. Exp. Psychol. Gen.* (submitted).
17. Anderson, A. K. & Phelps, E. A. Expression without recognition: Contributions of the human amygdala to emotional communication. *Psychol. Sci.* **11**, 106–111 (2000).
18. Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. R. Fear and the human amygdala. *J. Neurosci.* **15**, 5879–5891 (1995).
19. Strange, B. A., Henson, R. N. A., Friston, K. J. & Dolan, R. J. Brain mechanisms for detecting perceptual, semantic, and emotional deviance. *NeuroImage* **12**, 425–433 (2000).
20. Anderson, A. K., Spencer, D. D., Fulbright, R. K. & Phelps, E. A. Contribution of the anteromedial temporal lobes to the evaluation of facial emotion. *Neuropsychology* **14**, 526–536 (2000).
21. Adolphs, R., Damasio, H., Tranel, D., Cooper, G. & Damasio, A. R. A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *J. Neurosci.* **20**, 2683–2690 (2000).
22. Funayama, E. S., Grillon, C. G., Davis, M. & Phelps, E. A. A double dissociation in the affective modulation of startle in humans: Effects of unilateral temporal lobectomy. *J. Cogn. Neurosci.* (in the press).
23. Kapp, B. S., Wilson, A., Pascoe, J. P., Supple, W. & Whalen, P. J. In *Neurocomputation and Learning: Foundations of Adaptive Networks* (eds Gabriel, M. & Moore, J.) 53–90 (MIT Press, Cambridge, Massachusetts, 1990).
24. Whalen, P. J. Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Curr. Dir. Psychol. Sci.* **7**, 177–188 (1998).
25. Morris, J. S. *et al.* A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain* **121**, 47–57 (1998).
26. Morris, J. S., Friston, K. J. & Dolan, R. J. Experience-dependent modulation of tonotopic neural responses in human auditory cortex. *Proc. R. Soc. Lond. B* **265**, 649–657 (1998).
27. Isenberg, N. *et al.* Linguistic threat activates the human amygdala. *Proc. Natl Acad. Sci. USA* **96**, 10456–10459 (1999).
28. Weinberger, N. M. In *The Cognitive Neurosciences* (ed. Gazzaniga, M. S.) 1071–1089 (MIT Press, Cambridge, Massachusetts, 1995).
29. Adolphs, R. & Tranel, D. In *The Amygdala: A Functional Analysis* (ed. Aggleton, J. P.) 587–630 (Oxford Univ. Press, New York, 2000).
30. Spencer, D. D., Spencer, S. S., Mattson, R. H., Williamson, P. D. & Novelly, R. A. Access to the posterior medial temporal lobe structure in the surgical treatment of temporal lobe epilepsy. *Neurosurgery* **15**, 667–671 (1984).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Chun, E. De Rosa, K. O'Connor, K. Ochsner, I. Olson, M. Packard and N. Sobel for comments during the preparation of this manuscript, and D. Spencer for access to the patient samples. This work was supported by the James S. McDonnell Foundation (E.A.P.).

Correspondence and requests for materials should be addressed to A.K.A. (e-mail: adam.k.anderson@stanford.edu).

α-CaMKII-dependent plasticity in the cortex is required for permanent memory

Paul W. Frankland*, Cara O'Brien†, Masuo Ohno*, Alfredo Kirkwood† & Alcino J. Silva*

* Departments of Neurobiology, Psychiatry and Psychology, Brain Research Institute, University of California, Los Angeles, California 90095-1761, USA

† Department of Neuroscience and Mid/Brain Institute, Johns Hopkins University, Baltimore, Maryland 21218, USA

Cortical plasticity seems to be critical for the establishment of permanent memory traces^{1–3}. Little is known, however, about the molecular and cellular processes that support consolidation of memories in cortical networks^{4,5}. Here we show that mice heterozygous for a null mutation of α-calcium-calmodulin kinase II (α-CaMKII^{+/-}) show normal learning and memory 1–3 days after training in two hippocampus-dependent tasks. However, their memory is severely impaired at longer retention delays (10–50 days). Consistent with this, we found that α-CaMKII^{+/-} mice have impaired cortical, but not hippocampal, long-term potentiation. Our results represent a first step in unveiling the molecular and cellular mechanisms underlying the establishment of permanent memories, and they indicate that α-CaMKII may modulate the synaptic events required for the consolidation of memory traces in cortical networks.

A homozygous null mutation of α-CaMKII blocks hippocampal long-term potentiation (LTP) and learning^{6,7}, and disrupts experience-dependent plasticity in the sensory cortex without affecting cortical topography^{8,9}. In contrast, α-CaMKII heterozygous mice learn normally in hippocampus-dependent tasks under moderate-to-intensive training conditions⁷. We were, therefore, able to study the impact of reduced levels of α-CaMKII on the establishment of permanent memories. We trained α-CaMKII^{+/-} mutants and their wild-type littermate controls in contextual fear conditioning. In contextual fear conditioning, an association is formed between a distinctive context and an aversive event that takes place in that context. When placed back in the context, mice exhibit a range of conditioned fear responses, including freezing (the absence of all but respiratory movement). This is a particularly robust and long-lasting form of learning that is dependent on the hippocampus^{10,11}. Importantly, post-training lesions of the hippocampus disrupt recent, but not remote, contextual fear memories¹⁰.

After training with three shocks, α-CaMKII^{+/-} mutants and wild-type mice were placed back in the context at different times (1, 3, 10, 17 or 50 days), and freezing behaviour was recorded using automated procedures¹². To avoid the confounding effects of extinction, separate groups of mice were tested at each retention delay. Whereas wild-type mice showed stable contextual freezing at all test delays, contextual freezing declined sharply at longer test delays in α-CaMKII^{+/-} mutants (genotype × delay interaction, $F(4, 102) = 17.6$, $P < 0.01$) (Fig. 1a).

During training, α-CaMKII^{+/-} mutants were around 20% more active than the wild-type mice before shock delivery ($F(1, 102) = 14.8$, $P < 0.05$) (data not shown). We therefore assessed contextual memory with a different measure that is not affected by differences in basal activity. In the same mice, we calculated activity suppression by comparing activity levels before a shock delivery during training with activity levels during the test, thus normalizing for initial differences in activity¹². Although wild-type and α-CaMKII^{+/-} mice showed indistinguishable levels of activity suppression one day after training, at longer delays, activity

suppression was impaired in α -CaMKII^{+/-} mutants but not in wild-type controls (genotype $\times$ delay interaction, $F(4, 102) = 28.5$, $P < 0.05$) (Fig. 1b). Post-hoc analyses confirmed that levels of activity suppression were greater in wild-type controls than in α -CaMKII^{+/-} mutants at all retention delays except at day one ($P < 0.05$).

We next tested whether intense training could compensate for the memory deficits of the α -CaMKII^{+/-} mutants. We trained mice with eight shocks and tested separate groups 1, 3 or 17 days after training. Consistent with our previous experiment, freezing declined in mutants, but not in wild-type controls, at longer test delays (genotype $\times$ delay interaction, $F(2, 41) = 9.1$, $P < 0.05$) (Fig. 1c). Post-hoc analyses confirmed that freezing was lower in α -CaMKII^{+/-} mutants only in the 3- and 17-day tests ($P < 0.05$). A similar pattern was observed when activity suppression was used as an index of contextual fear (Fig. 1d). Activity suppression was impaired in α -CaMKII^{+/-} mutants, but not wild-type mice, at longer test delays (genotype $\times$ delay interaction, $F(2, 41) = 11.0$, $P < 0.05$). Therefore, after either moderate (three shocks) or intense (eight shocks) training, memory loss in α -CaMKII^{+/-} mutants is specific to longer retention delays. Protein-synthesis blockers impair contextual memory 3–5 hours after training¹³. The similar levels of contextual fear in the mutant and wild-type mice 24 h after training indicate that protein synthesis-dependent processes underlying the initial formation of hippocampus-dependent memories are unaffected in these mutant mice. Notably, the time course of memory decay in α -CaMKII^{+/-} mutants corresponds to the shift from hippocampus-dependent to hippocampus-independent retrieval inferred from imaging and lesion studies in mice and rats^{10,14,15}. For example, the genetic disruption of NMDA (N-methyl-D-aspartate) receptors in the CA1 region of the hippocampus during the 0–14-day period after training blocks contextual memories in mice. In contrast, similar disruption three weeks after training leaves these memories intact¹⁵.

Under less intensive training conditions, α -CaMKII^{+/-} mutants exhibit learning impairments that are affected by genetic background (data not shown)⁷. Thus, despite equivalent levels of freezing in the retention test at day one, the initial underlying memory trace might be weaker in α -CaMKII^{+/-} mutants, and hence less durable. To address this, we compared contextual memory in wild-type mice after weak training (one shock) with contextual memory in α -CaMKII^{+/-} mutants that had been trained intensively

(eight shocks). Wild-type mice, trained at the same time as the α -CaMKII^{+/-} mutants, were tested 1, 3 or 17 days after training. WT mice exhibited much lower freezing levels than α -CaMKII^{+/-} mutants at the one-day retention delay. However, unlike the mutants, their levels of freezing were similar at all retention intervals (genotype $\times$ delay interaction, $F(2, 54) = 13.7$, $P < 0.05$) (Fig. 1e), indicating that lower initial levels of freezing do not necessarily result in unstable memory. Post-hoc analyses revealed that α -CaMKII^{+/-} mutants froze more than wild-type controls one day after training, but significantly less than wild-type controls 17 days after training ($P < 0.05$).

To assess the severity of memory loss in mutants at long delays, we compared the acquisition of contextual fear conditioning in naive (non-trained) α -CaMKII^{+/-} mutants and α -CaMKII^{+/-} mutants that had been trained 30 days before with three shocks. Both naive and previously trained mutants showed similar rates of acquisition of contextual freezing when trained with one shock per day over four days (Fig. 1f); this was supported by an absence of a condition (trained versus naive) $\times$ day (1–4) interaction ($F(3, 30) = 0.20$, $P = 0.89$). The absence of accelerated acquisition (or behavioural savings) in the previously trained mutants is consistent with the hypothesis that memory loss at long retention delays is substantial.

To test whether our initial findings applied to another form of hippocampal learning, we trained α -CaMKII^{+/-} mutants in the Morris water maze¹⁶. In this spatial learning task, mice learn to navigate to a submerged platform by using extra-maze cues. Mice were trained with 12 trials per day over 5 days, as under these conditions α -CaMKII^{+/-} mutants performed this task as well as wild-type controls⁷. At the completion of training, spatial learning was assessed in a 60-s probe test where the platform was removed from the pool (Fig. 2a–c). To assess spatial memory at different retention delays, separate groups of mice were tested 3, 10 or 17 days after completing the training (Fig. 2d–f). No other training was given between the probe test at the end of training and the memory probe test. In probe tests at the completion of training, wild-type and α -CaMKII^{+/-} mice spent equivalent amounts of time in the target quadrant of the pool ($F(1, 49) = 0.93$, $P = 0.34$). Similarly, α -CaMKII^{+/-} mutants exhibited normal spatial memory when tested three days after completing the training ($F(1, 16) = 0.37$, $P = 0.55$). Therefore, consistent with the fear conditioning data, α -CaMKII^{+/-} mutants showed normal acquisition and normal

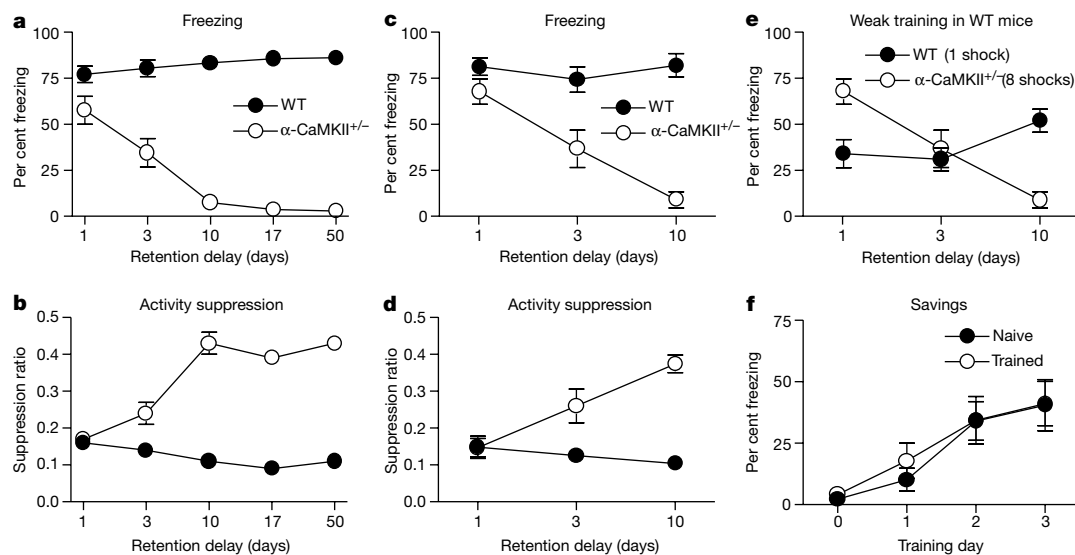


Figure 1 Contextual memory in α -CaMKII^{+/-} mutants. **a, b**, Freezing and activity suppression for wild-type (WT) and α -CaMKII^{+/-} mice trained with three shocks and tested at different retention delays. **c, d**, Freezing and activity suppression for WT and α -CaMKII^{+/-} mice trained with eight shocks and tested at different retention delays. **e**,

Freezing at different retention delays in WT mice after weak training (one shock) and α -CaMKII^{+/-} mutants after intensive training (eight shocks). **f**, Acquisition of contextual freezing in previously trained and naive α -CaMKII^{+/-} mutants.

memory at short retention delays, indicating that motivation, vision and motor skills required for the task were not affected. However, α -CaMKII^{+/-} mutants spent significantly less time searching the target quadrant when tested 10 days ($F(1, 15) = 15.2$, $P < 0.05$) and 17 days ($F(1, 18) = 10.1$, $P < 0.05$) after the completion of training. Notably, there was a significant genotype $\times$ delay interaction ($F(2, 49) = 3.23$, $P < 0.05$), reflecting poorer spatial memory in α -CaMKII^{+/-} mutants only at long retention delays (Fig. 2g). Retrograde amnesia after hippocampal lesions is typically not graded in water maze studies¹⁷, possibly because the hippocampus is required for some aspects of performance during navigation, such as path integration^{18–20}. However, under the training conditions used, α -CaMKII^{+/-} mutants learned the water maze normally, indicating that the processing required for performance of the task, such as path integration, is not affected under these conditions in these mice. Rather, the poor performance at long retention delays is likely to reflect loss of spatial memory.

These data show that α -CaMKII^{+/-} mutants have deficits in memory at a time when memories become more dependent on cortical function, and less on the hippocampus^{10,14}. Whereas hippocampal plasticity is required for initial learning and memory formation, cortical plasticity is thought to be critical for the

establishment of permanent memory traces^{1–3,14}. Therefore, hippocampal and cortical synaptic plasticity may be differentially affected in α -CaMKII^{+/-} mutants^{7,21}. To investigate this possibility, we examined LTP induced by five θ -burst stimulations (TBS) in the hippocampus and cortex in the same slices from α -CaMKII^{+/-} mutants and wild-type mice (Fig. 3). This strong induction protocol saturates LTP in the cortex (A.K., unpublished observations). Memory is likely to be widely distributed within cortical networks³, so we chose to study the visual cortex as a representative of primary sensory cortex and the temporal cortex as a representative of associational cortex. Stimulation of Schaffer collaterals produced stable LTP in the CA1 region of the hippocampus of wild-type ($t(10) = 7.17$, $P < 0.05$) and of α -CaMKII^{+/-} mice ($t(8) = 4.54$, $P < 0.05$), and the magnitude of LTP did not differ between the two ($F(1, 18) = 1.83$, $P = 0.19$). In addition, protein-synthesis-

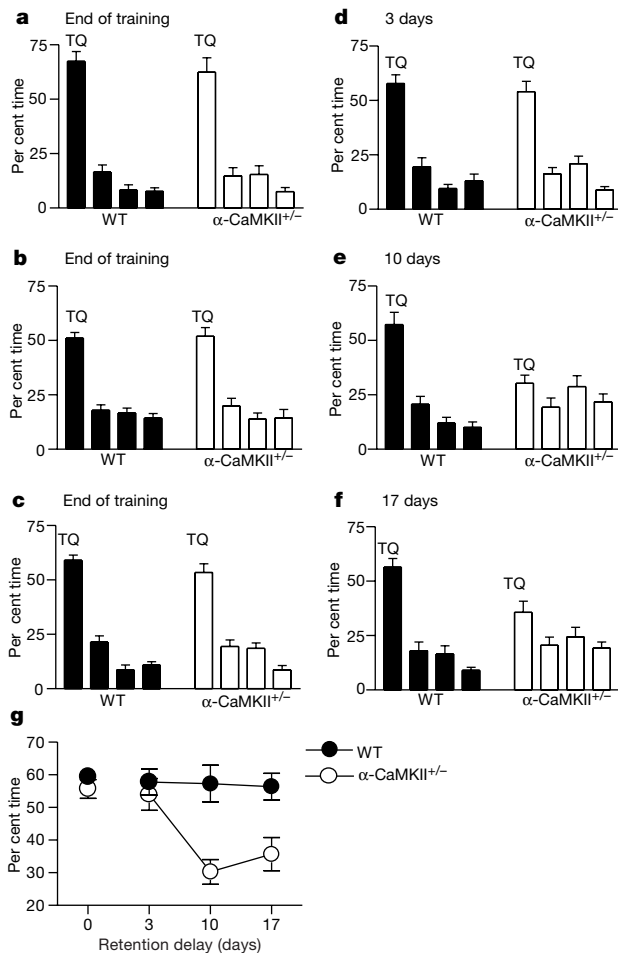


Figure 2 Spatial memory in α -CaMKII^{+/-} mutants. Mice were given a probe test at the end of training (a–c) and then tested 3, 10 or 17 days after training (d–f). The per cent time searching the target quadrant (TQ) of the pool in which the platform was located during training versus each of the other three quadrants is plotted for WT and α -CaMKII^{+/-} mutants. After training, WT mice and α -CaMKII^{+/-} mutants searched selectively in the TQ. Spatial memory in α -CaMKII^{+/-} mutants was normal three days after training, but impaired in the 10- or 17-day retention tests. g, Summary of results of probe tests for WT and α -CaMKII^{+/-} mice. The per cent time searching the TQ is plotted for each of the probe tests at different times after the completion of training.

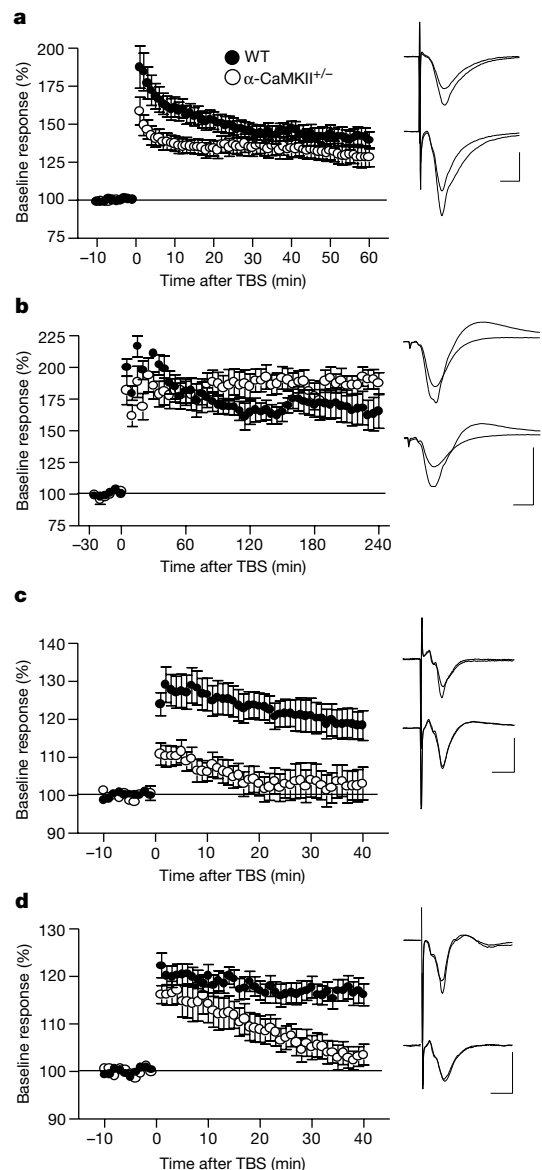


Figure 3 Reduced cortical LTP in α -CaMKII^{+/-} mutants. The graphs depict the time course of the changes in magnitude of synaptic responses following TBS for recordings in the CA1 region of the hippocampus (a, LTP; b, late LTP), visual cortex (c), and temporal cortex (d) for slices from WT and α -CaMKII^{+/-} mice. Examples of LTP are shown at the right of each graph (WT, upper; α -CaMKII^{+/-}, lower). These superimposed traces are the average of four consecutive responses recorded 1 min before TBS and at the end of the experiment (60 min for CA1 LTP, 240 min for CA1 late LTP, 40 min for visual and temporal cortex LTP). Calibration bars: 5 ms, 1 mV.

dependent late LTP, induced by three trains of ten TBS, was normal in α -CaMKII^{+/−} mutants ($F(1, 11) = 2.87$, $P > 0.05$). Post-tetanic potentiation was reduced in α -CaMKII^{+/−} mutants⁷. This is consistent with our behavioural data showing that memory is normal in α -CaMKII^{+/−} mutants up to three days after training. In contrast, stimulation of layer IV failed to produce LTP in visual cortex ($t(10) = 1.30$, $P > 0.05$) and temporal cortex ($t(22) = 1.82$, $P > 0.05$) in α -CaMKII^{+/−} mutants. Similar stimulation produced stable LTP in visual cortex ($t(18) = 4.23$, $P < 0.05$) and temporal cortex ($t(21) = 7.00$, $P < 0.05$) in wild-type mice, and these levels of potentiation were significantly higher than those in α -CaMKII^{+/−} mutants (visual cortex: $F(1, 28) = 5.87$, $P < 0.05$; temporal cortex: $F(1, 43) = 15.3$, $P < 0.05$). Consistent with previous data, these results indicate that LTP impairments in α -CaMKII^{+/−} mutants are restricted to the cortex^{7,21}. Both hippocampal and cortical LTP are impaired in α -CaMKII homozygous mice^{7,21}. The α -CaMKII protein is expressed more abundantly in the hippocampus than in the cortex, so it is likely that the loss of half the normal levels of this kinase affects cortical plasticity more severely.

In humans and animals, damage limited to the hippocampal formation produces a severe amnesia for recently acquired information that is declarative or episodic in nature, but spares memory for similar remotely acquired information^{10,22–24}. The opposite pattern is observed in cases of semantic dementia associated with atrophy of the temporal cortex, which leads to impairments of remote, rather than recent, memory². Consistent with this, retrieval of recently acquired spatial information is associated with activation of the hippocampus, whereas retrieval of remotely acquired spatial information is associated with activation of the cortex in mice¹⁴. This has led to the idea that storage and retrieval of recent memories is dependent on hippocampal networks, whereas storage and retrieval of remote memories depends on cortical^{1–3} or cortical–hippocampal²⁵ networks. Although hippocampal lesions preferentially disrupt recent but not remote (or permanent) memories, we found that a functional disruption of cortical plasticity preferentially disrupts the establishment of permanent memories. Furthermore, the time course of the memory decay in the α -CaMKII^{+/−} mutants corresponds to that predicted by imaging and lesion studies in mice and rats^{10,14,15,26}. In addition, we found that mice that are heterozygous for a point mutation that substitutes threonine 286 for alanine in α -CaMKII (α -CaMKII^{T286A/+})²⁷ exhibit normal contextual and spatial memory for up to 50 days (data not shown). Because cortical plasticity is not impaired in these mice²⁸, this finding strengthens the conclusion that normal cortical plasticity is required for the establishment of permanent memory traces. The consolidation of memory traces in cortical networks is likely to depend on the recurrent activation of cortical traces during, for example, sleep³. Our results indicate that repetitive LTP-like events, which are dependent on α -CaMKII, are required for the consolidation of memory traces in cortical networks, and they represent a first step in unveiling the molecular and cellular mechanisms underlying cortical memory consolidation. They further indicate that the molecular processes underlying the establishment of permanent cortical memory traces may share some of the same mechanisms as those underlying initial encoding of hippocampal memories. □

Methods

Mice

The α -CaMKII^{+/−} mice used in the fear conditioning and electrophysiological experiments were the F₁ progeny from a cross between α -CaMKII^{+/−} mice in the C57B1/6 background (> 99%) and 129/SvEms mice. For the water maze studies, mice were tested from two different genetic backgrounds with similar findings. The two backgrounds tested included mice that were the offspring from a cross between α -CaMKII^{+/−} mice in the C57B1/6 background (> 99%) and 129/SvEms mice, and mice resulting from a cross between chimaeras having the α -CaMKII mutation in the 129/Ola background and C57B1/6 mice. The C57B1/6 background exacerbates the effects of the α -CaMKII heterozygous mutation (data not shown). Mice were housed in groups and maintained on a 12 h light/dark cycle.

Mice (at least eight weeks old) were tested during the light phase of the cycle. All behavioural and electrophysiological experiments were conducted blind to the genotype.

Contextual conditioning

In each of the experiments, mice were placed in the conditioning chamber and presented with a series of unignalled foot-shocks (2 s duration, 0.5 mA, 1 min apart). During testing, mice were placed back in the conditioning chamber and freezing and activity were recorded using automated procedures described previously¹². Activity suppression ratios were calculated by comparing activity levels during the first 2 min of training with the first 2 min of testing according to the formula: $(\text{activity}_{\text{test}}/(\text{activity}_{\text{train}} + \text{activity}_{\text{test}}))$. In the first experiment, mice were trained with 3 foot-shocks and tested 1 (wild-type, $n = 8$; α -CaMKII^{+/−}, $n = 9$), 3 (wild-type, $n = 14$; α -CaMKII^{+/−}, $n = 12$), 10 (wild-type, $n = 16$; α -CaMKII^{+/−}, $n = 10$), 17 (wild-type, $n = 14$; α -CaMKII^{+/−}, $n = 11$) or 50 (wild-type, $n = 9$; α -CaMKII^{+/−}, $n = 9$) days after training. In the second experiment, mice were trained with 8 foot-shocks and tested 1 (wild-type, $n = 9$; α -CaMKII^{+/−}, $n = 6$), 3 (wild-type, $n = 7$; α -CaMKII^{+/−}, $n = 9$) or 17 (wild-type, $n = 8$; α -CaMKII^{+/−}, $n = 9$) days after training. In the third experiment, wild-type mice were trained with 1 foot-shock (after 3 min of a 5-min session) and tested 1 ($n = 9$), 3 ($n = 14$) or 17 ($n = 13$) days after training. In the fourth experiment, acquisition of contextual fear conditioning was compared in naive mutants and mutants that had been trained with 3 shocks 30 days earlier. During acquisition, naive ($n = 6$) and trained ($n = 6$) mutants were trained with a single foot-shock (delivered after 3 min of a 5-min session) over four consecutive days. Freezing was measured before the shock was delivered on each day.

Water maze

The water maze apparatus and procedures have been described²⁹. Mice were trained with 12 trials per day for 5 days (3 blocks of 4 trials; 1 min inter-trial intervals, 1 h inter-block intervals). Spatial learning and memory were assessed in probe tests with the platform removed from the pool. All mice were given a probe test at the end of training, and a memory probe test 3 (wild-type, $n = 10$; α -CaMKII^{+/−}, $n = 8$), 10 (wild-type, $n = 9$; α -CaMKII^{+/−}, $n = 8$) or 17 (wild-type, $n = 10$; α -CaMKII^{+/−}, $n = 10$) days later.

Electrophysiology

Coronal brain slices from 6–8-week-old mice containing visual cortex, temporal cortex and the CA1 region of the hippocampus were prepared as described²¹. Slices were left to recover for at least 1 h before recording in oxygenated (95% O₂ and 5% CO₂) warm (30 °C) artificial cerebrospinal fluid (ACSF) containing (in mM): 124 NaCl, 5 KCl, 1.25 NaH₂PO₄, 1 MgCl₂, 2 CaCl₂, 26 NaHCO₃, 10 dextrose. The experiments were performed in an interface chamber with an atmosphere of humidified 95% O₂ and 5% CO₂, and superfused with 30 °C ACSF at a rate of 1 ml min^{−1}. Field potentials (FP) in layer III evoked by layer IV stimulation, and CA1 FP evoked by Schaffer collateral stimulation, were measured as previously described³⁰. Responses were quantified either as the initial slope of the FP in CA1, or as the amplitude of the FP in cortex. The amplitude, rather than slope, of the FP in layer III was used as a measure of the evoked population excitatory synaptic current, because the initial slope was contaminated by the presynaptic volley. LTP was induced with TBS, which consisted of five brief bursts (four pulses at 100 Hz) of stimuli delivered every 200 ms. Average responses ($\pm$ s.e.m.) were expressed as percentage of pre-TBS baseline responses (at least 10 min of stable responses). LTP induction was assessed with paired t -tests by comparing responses recorded before and 60 min (CA1 LTP) or 40 min (cortical LTP) after TBS. To examine late LTP in CA1, transverse hippocampal slices were maintained in a submerged chamber as previously described²⁷. Late LTP was induced by 3 trains of 10 TBS spaced 10 min apart. To determine whether the magnitude of LTP differed between wild-type and α -CaMKII^{+/−} mice, responses from the last 10-min block of recordings (51–60 min for CA1 LTP; 31–40 min for cortical LTP) or the last 30-min block of recordings (210–240 min for CA1 late LTP) were compared using analysis of variance.

Received 2 November 2000; accepted 12 March 2001.

- Squire, L. R. & Alvarez, P. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Curr. Opin. Neurobiol.* **5**, 169–177 (1995).
- Graham, K. S., Patterson, K. & Hodges, J. R. Episodic memory: new insights from the study of semantic dementia. *Curr. Opin. Neurobiol.* **9**, 245–250 (1999).
- McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
- Dudai, Y. Consolidation: fragility on the road to the engram. *Neuron* **17**, 367–370 (1996).
- McGaugh, J. L. Memory—a century of consolidation. *Science* **287**, 248–251 (2000).
- Silva, A. J., Paylor, R., Wehner, J. M. & Tonegawa, S. Impaired spatial learning in alpha-calcium-calmodulin kinase II mutant mice. *Science* **257**, 206–211 (1992).
- Silva, A. J. *et al.* Impaired learning in mice with abnormal short-lived plasticity. *Curr. Biol.* **6**, 1509–1518 (1996).
- Glazewski, S., Chen, C. M., Silva, A. & Fox, K. Requirement for alpha-CaMKII in experience-dependent plasticity of the barrel cortex. *Science* **272**, 421–423 (1996).
- Gordon, J. A., Cioffi, D., Silva, A. J. & Stryker, M. P. Deficient plasticity in the primary visual cortex of alpha-calcium/calmodulin-dependent protein kinase II mutant mice. *Neuron* **17**, 491–496 (1996).
- Kim, J. J. & Fanselow, M. S. Modality-specific retrograde amnesia of fear. *Science* **256**, 675–677 (1992).
- Frankland, P. W., Cestari, V., Filipkowski, R. K., McDonald, R. J. & Silva, A. J. The dorsal hippocampus is essential for context discrimination but not for contextual conditioning. *Behav. Neurosci.* **112**, 863–874 (1998).
- Anagnostaras, S. G., Josselyn, S. A., Frankland, P. W. & Silva, A. J. Computer-assisted behavioral assessment of Pavlovian fear conditioning in mice. *Learn. Mem.* **7**, 58–72 (2000).

13. Bourtochouladze, R. *et al.* Different training procedures recruit either one or two critical periods for contextual memory consolidation, each of which requires protein synthesis and PKA. *Learn. Mem.* **5**, 365–374 (1998).
14. Bontempi, B., Laurent-Demir, C., Destrade, C. & Jaffard, R. Time-dependent reorganization of brain circuitry underlying long-term memory storage. *Nature* **400**, 671–675 (1999).
15. Shimizu, E., Tang, Y. P., Rampon, C. & Tsien, J. Z. NMDA receptor-dependent synaptic reinforcement as a crucial process for memory consolidation. *Science* **290**, 1170–1174 (2000).
16. Morris, R. G. M., Garrud, P., Rawlins, J. N. P. & O'Keefe, J. Place navigation impaired in rats with hippocampal lesions. *Nature* **297**, 681–683 (1982).
17. Mumby, D. G., Astur, R. S., Weisend, M. P. & Sutherland, R. J. Retrograde amnesia and selective damage to the hippocampal formation: memory for places and object discriminations. *Behav. Brain Res.* **106**, 97–107 (1999).
18. Riedel, G. *et al.* Reversible neural inactivation reveals hippocampal participation in several memory processes. *Nature Neurosci.* **2**, 898–905 (1999).
19. Knowlton, B. J. & Fanselow, M. S. The hippocampus, consolidation and on-line memory. *Curr. Opin. Neurobiol.* **8**, 293–296 (1998).
20. Whishaw, I. Q., McKenna, J. E. & Maaswinkel, H. Hippocampal lesions and path integration. *Curr. Opin. Neurobiol.* **7**, 228–234 (1997).
21. Kirkwood, A., Silva, A. & Bear, M. F. Age-dependent decrease of synaptic plasticity in the neocortex of alphaCaMKII mutant mice. *Proc. Natl Acad. Sci. USA* **94**, 3380–3383 (1997).
22. Scoville, W. B. & Milner, B. Loss of recent memory after hippocampal bilateral lesions. *J. Neurol. Neurosurg. Psychiatry* **20**, 11–12 (1957).
23. Rempel-Clower, N. L., Zola, S. M., Squire, L. R. & Amaral, D. G. Three cases of enduring memory impairment after bilateral damage limited to the hippocampal formation. *J. Neurosci.* **16**, 5233–5255 (1996).
24. Zola-Morgan, S. M. & Squire, L. R. The primate hippocampal formation: evidence for a time-limited role in memory storage. *Science* **250**, 288–290 (1990).
25. Nadel, L. & Moscovitch, M. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).
26. Cho, Y. H., Beracochea, D. & Jaffard, R. Extended temporal gradient for the retrograde and anterograde amnesia produced by ibotenate entorhinal cortex lesions in mice. *J. Neurosci.* **13**, 1759–1766 (1993).
27. Giese, K. P., Fedorov, N. B., Filipkowski, R. K. & Silva, A. J. Autophosphorylation at Thr286 of the α -calcium-calmodulin kinase II in LTP and learning. *Science* **279**, 870–873 (1998).
28. Glazewski, S., Giese, K., Silva, A. & Fox, K. The role of α -CaMKII autophosphorylation in neocortical experience-dependent plasticity. *Nature Neurosci.* **3**, 911–918 (2000).
29. Bourtochouladze, R. *et al.* Deficient long-term memory in mice with a targeted mutation of the cAMP-responsive element-binding protein. *Cell* **79**, 59–68 (1994).
30. Kirkwood, A. & Bear, M. F. Hebbian synapses in visual cortex. *J. Neurosci.* **14**, 1634–1645 (1994).

Acknowledgements

We thank S. Anagnostaras, R. Costa, S. Josselyn, S. Köhler, S. Kushner, G. Murphy and K. Nader for comments on the paper. These studies were supported by grants from NIH, National Eye Institute (A.K.), FRAXA Research Foundation (P.W.F.), and the Whitehall, Beckman, Merck, Klingenstein and McKnight Foundations (A.J.S.).

Correspondence and requests for materials should be addressed to A.J.S. (e-mail: silvaa@ucla.edu).

A new role for cryptochrome in a *Drosophila* circadian oscillator

Balaji Krishnan[†], Joel D. Levine[†], M. Kathlea S. Lynch[‡], Harold B. Dowse[§], Pablo Funes[‡], Jeffrey C. Hall[‡], Paul E. Hardin[†] & Stuart E. Dryer[†]

[†] Department of Biology and Biochemistry and Biological Clocks Program, University of Houston, Houston, Texas 77204, USA

[‡] Department of Biology and NSF Center for Biological Timing, Brandeis University, Waltham, Massachusetts 02454, USA

[§] Department of Biological Sciences, University of Maine, Orono, Maine 04469, USA

* These authors contributed equally to this work

Cryptochromes are flavin/pterin-containing proteins that are involved in circadian clock function in *Drosophila* and mice. In mice, the cryptochromes Cry1 and Cry2 are integral components of the circadian oscillator within the brain^{1–6} and contribute to circadian photoreception in the retina⁷. In *Drosophila*, cryptochrome (CRY) acts as a photoreceptor that mediates light input to circadian oscillators in both brain and peripheral tissue^{8–12}. A *Drosophila* cry mutant, *cry^b*, leaves circadian oscillator function

intact in central circadian pacemaker neurons but renders peripheral circadian oscillators largely arrhythmic. Although this arrhythmicity could be caused by a loss of light entrainment, it is also consistent with a role for CRY in the oscillator. A peripheral oscillator drives circadian olfactory responses in *Drosophila* antennae¹³. Here we show that CRY contributes to oscillator function and physiological output rhythms in the antenna during and after entrainment to light–dark cycles and after photic input is eliminated by entraining flies to temperature cycles. These results demonstrate a photoreceptor-independent role for CRY in the periphery and imply fundamental differences between central and peripheral oscillator mechanisms in *Drosophila*.

We initially established that the *cry^b* mutant severely reduces or eliminates circadian rhythms in olfactory responses and molecular rhythms in oscillator components during and after entrainment to light–dark cycles. Electroantennogram (EAG) responses to the food odorant ethyl acetate were measured in wild-type and *cry^b* flies under 12 h:12 h light–dark cycles (LD) and on the second day of constant darkness (DD) after LD entrainment. In *cry^b* mutants, we observed EAG responses of similar magnitude throughout the day except for a weak but statistically significant peak at Zeitgeber time 17 (5 h after lights off; Fig. 1a), indicating that the *cry^b* allele is nearly a loss-of-function mutant. This small peak was lost when mutant antennae were recorded in DD (Fig. 1b). Thus, *cry⁺* is required for a

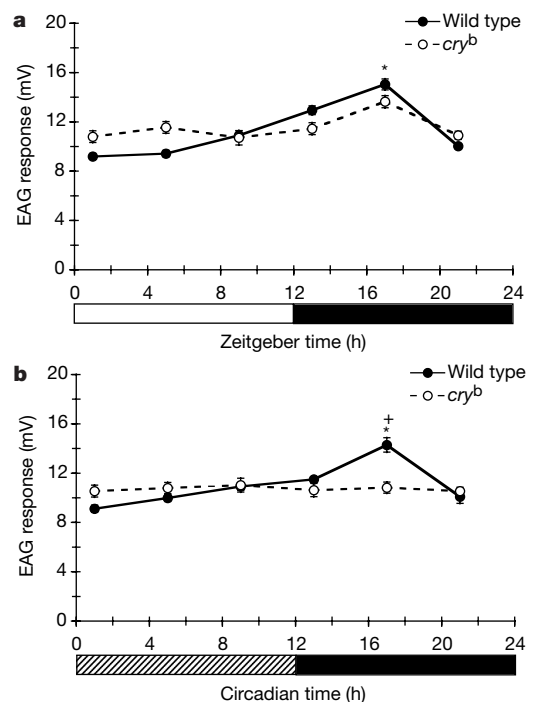


Figure 1 Olfactory responses of wild-type and *cry^b* mutant *Drosophila* under LD and DD conditions. Each point represents mean odorant-evoked EAG responses. Error bars denote s.e.m. **a**, Responses on the fourth day of LD 12 h:12 h. $n = 48$ flies per point. White bar, lights on; black bar, lights off. Overall effects of time of day and interaction were significant ($P < 0.05$) by two-way analysis of variance (ANOVA), but effects of genotype were not. Asterisk, significant ($P < 0.005$) difference in wild-type responses at Zeitgeber time (ZT)17 compared to all other times of day. The *cry^b* response at ZT17 was significantly ($P < 0.0001$) different from that at ZT1, but not other times of day. **b**, Responses on the second day of DD, indicating lack of rhythmicity in the mutant. $n = 24$ flies per point. Hatched bar, subjective day; black bar, subjective night. Cross, significant ($P < 0.005$) difference between wild type at ZT17 and *cry^b* at all times of day. Overall effects of time of day and interaction were significant ($P < 0.000001$), but effects of genotype were not. Responses of *cry^b* in LD and DD were not different ($P > 0.1$) at any time of day by three-way ANOVA.

physiological clock output from the antenna in constant conditions.

EAG experiments measure a physiological index of circadian clock function that is downstream from the core oscillator itself. However, a local circadian oscillator in the antenna can be monitored directly using bioluminescence rhythms in transgenic *per-luc* or *tim-luc* reporter strains^{14–16}. To determine whether *cry^b* affects oscillator function in antennae, we monitored bioluminescence under LD and DD conditions in isolated pairs of antennae from

wild-type and *cry^b* flies bearing *per-luc* or *tim-luc* reporter genes (Fig. 2). Robust rhythms in bioluminescence were seen under both LD and DD conditions in most (84% in LD and 77% in DD) wild-type antennae (see Table 1). In contrast, only 26% of *cry^b* specimens in LD showed bioluminescence rhythms, and these rhythms were much weaker than in *cry⁺* (Table 1). The *cry^b* mutation also increased the overall abundance of *luc*-reported levels of clock-gene products in the antenna (see Supplementary Information).

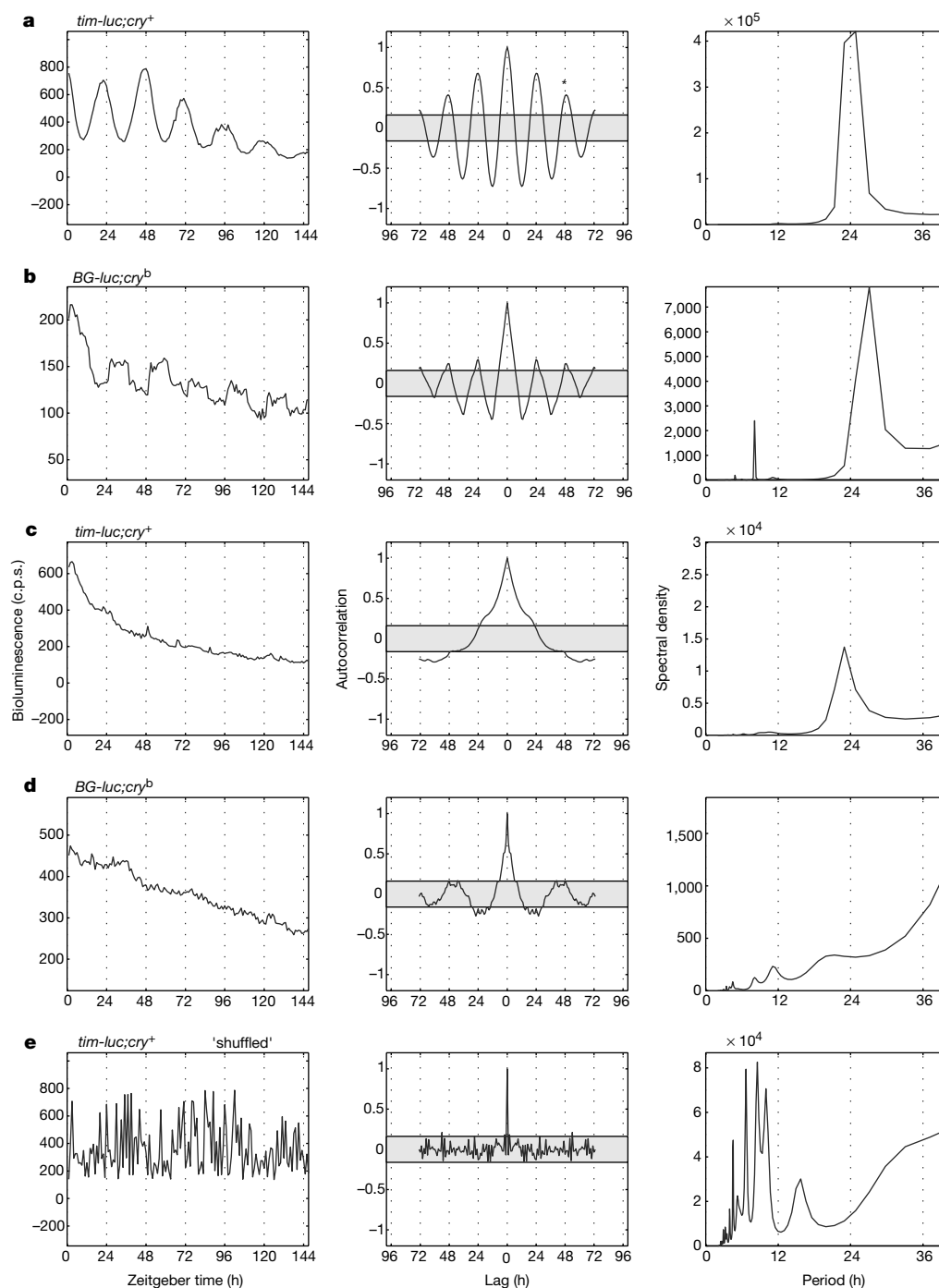


Figure 2 Examples of clock-gene-controlled luciferase (*luc*) activity rhythms from *Drosophila* antennae during LD. Left, bioluminescence plotted over time (Zeitgeber time 0–12, 24–36 h and so on represent lights on; 12–24, 36–48 h and so on, lights off). Middle, correlograms applied to assess the extent of any periodicity in the data. Shaded areas, 95% confidence intervals. Right, spectral power plotted as a function of period (MESA). **a** and **b** exemplify rhythmic samples; **c**–**e** were not significantly periodic. **a**, *Luc* activity driven by a 5′-flanking region of *tim* in wild-type antennae. Five complete cycles

are evident in the raw data plot. The correlogram shows significant rhythmicity. Asterisk, peak used for the rhythmicity index (in this case 0.4, see Table 1). MESA produced a peak at around 24 h. **b**, *per-luc* expression in a *cry^b* antennal pair. Rhythmicity is significant and has an index of 0.28. **c**, Arrhythmic *tim-luc; cry⁺* specimen. Arrhythmicity is based on the correlogram. **d**, Arrhythmic *per-luc* expression in a *cry^b* specimen. **e**, Data set from **a** shuffled randomly with respect to time (see Methods and Table 1).

This type of result—an effect on expression of other clock genes in addition to flattening the cycles of their product fluctuations—often occurs under the influence of a clock mutation (for example, refs 1–6, 15, 16). Rhythmicity was detected in only 12% of *cry^b* antennae when bioluminescence was measured in DD (Table 1). Together with the loss of EAG rhythmicity induced by *cry^b* in constant darkness, these molecular effects indicate that CRY may function as part of the antennal oscillator. If CRY's role were limited to photoreception, the EAG rhythm would free-run in LD and DD, similar to the effects of the *cry^b* mutation on adult locomotor activity⁸. However, it is possible that CRY functions as a photoreceptor earlier in development when it might be required to initiate (through entrainment) the functioning of peripheral clocks.

To distinguish between these alternatives, we entrained wild-type and *cry^b* flies in 12 h 18 °C: 12 h 27 °C temperature cycles in constant darkness, then tested their antennae for EAG responses. The animals used in these experiments were never exposed to visible (<600 nm) light. Temperature cycles can entrain locomotor activity rhythms and oscillations in the clock proteins PER and TIM^{8,17}; the phases of these biological and molecular rhythms are advanced relative to the behavioural or protein-abundance peaks observed in LD during constant temperature^{8,17}. During temperature cycles, wild-type flies showed a robust rhythm in EAG responses with a difference between daily peak and nadir of around 4 mV (Fig. 3). The peak in EAG responses occurred during the warm phase of the temperature cycle, compared with the olfactory-response peak during the dark phase of an LD cycle (Fig. 1). In contrast, *cry^b* flies showed non-rhythmic EAG responses. To ensure that the EAG rhythms in wild-type flies were not simply driven by temperature cycles, we measured EAG responses at constant low temperature (18 °C) and at constant high temperature (27 °C) after entrainment to 12 h:12 h temperature cycles. Wild-type flies exhibited rhythms in EAG responses at both constant temperatures, with a peak in the subjective high temperature (Fig. 3). In contrast, *cry^b* flies showed no EAG rhythms at either constant low or constant high temperature (Fig. 3). Because this experimental design completely bypasses photic entrainment pathways, these results indicate that the circadian oscillator driving this antennal output has been rendered arrhythmic in *cry^b* flies.

To determine directly the state of the antennal oscillator in wild-type and *cry^b* flies after temperature entrainment, we measured *tim-luc* and *per-luc* bioluminescence in isolated antennae after entrainment to 12 h:12 h temperature cycles (Fig. 2, Table 1). Two protocols gave essentially equivalent results (see Methods), so we pooled the data (Table 1). In *per-luc* and *tim-luc* flies carrying a normal *cry* allele, bioluminescence rhythms were present in 40% of antennae at constant high temperature (27 °C). As with the EAG

rhythms, the phase of *tim-luc* cycling following temperature entrainment of *cry^b* antennae was altered compared with rhythms in LD-entrained animals (Table 1). For *cry^b* specimens monitored in the same temperature-entrainment experiments, 12% of the antennae exhibited bioluminescence rhythms (Table 1). The proportion of rhythmic mutant samples is significantly less than in controls, although the free-running rhythms in both genetic backgrounds were similarly weak (Table 1). The stronger rhythmicity observed for *cry^b* antennae in light:dark cycles—compared with results from all the constant-condition monitorings (Table 1)—indicates that cryptochrome may mediate circadian photoreception in this peripheral tissue as well as subserving clock functions.

The residual molecular rhythmicity observed in *cry^b* antennae

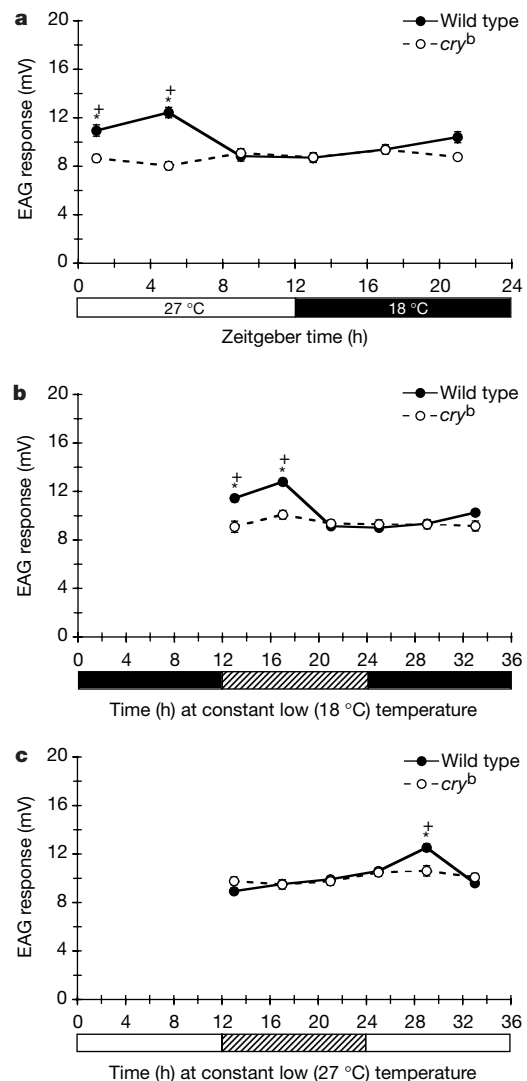


Figure 3 EAG responses of wild-type and *cry^b* flies during and after temperature entrainment. **a**, Responses during temperature entrainment. White bar, high (27 °C) temperature; black bar, low (18 °C) temperature. $n = 24$ flies per point. Overall effects of time of day, genotype and their interaction are significant ($P < 0.0000001$). Asterisk, significant ($P < 0.0005$) difference in wild-type flies at ZT17 versus ZT1 and ZT5. Crosses, significant ($P < 0.0001$) difference between wild-type and *cry^b* flies at the indicated times. **b**, Responses in constant low temperature. Black bar, subjective low temperature; hatched bar, subjective high temperature. $n = 48$ flies per point. Overall effects of time of day, genotype and their interaction are significant ($P < 0.0000001$). **c**, Responses in constant high temperature. White bar, subjective high temperature; hatched bar, subjective low temperature. $n = 48$ flies per point. Overall effects of time of day, genotype and their interaction are significant ($P < 0.0005$).

Table 1 Periodic components in *luc*-reported clock-gene expression from antennae

Condition	Genotype	n	No. rhythmic specimens			Rhythmicity index
			22–26 h	26–40 h	Shuffled	
LD 12:12*	<i>cry⁺</i>	56	47	0	0	0.31 ± 0.04
	<i>cry^b</i>	80	21	0	0	0.06 ± 0.04
DD†	<i>cry⁺</i>	61	47	0	0	−0.02 ± 0.02
	<i>cry^b</i>	68	5	3	0	−0.05 ± 0.06
Constant temperature‡	<i>cry⁺</i>	101	41	1	0	−0.02 ± 0.03
	<i>cry^b</i>	125	15	2	0	0.03 ± 0.01

n , number of antennal pairs analysed (see Supplementary Information). 22–26 h and 26–40 h are period ranges for significantly rhythmic specimens. Shuffled, the number of shuffled data sets that display circadian rhythmicity (see Methods and Fig. 2a, e). Rhythmicity index values ($\pm$ s.e.m.) are a measure of strength in rhythmic specimens. The proportions of 22–26 h rhythmic heads, bodies, wings and legs in LD were 68, 62, 79 and 81% in *cry⁺* and 0, 0, 41 and 4% in *cry^b*, respectively. Phase values for *cry⁺* tissues during temperature entrainment were ~3 h earlier compared to LD (see § in Supplementary Information).

* Rhythms in *cry⁺* and *cry^b* antennae are different (χ^2 , $P < 0.0001$). Rhythmicity index values for *cry⁺* were different from *cry^b* (t -test, $P < 0.0001$). Phase and *luc* expression levels are different in *cry⁺* and *cry^b* antennae (see Supplementary Information).

† The difference between rhythmic *cry⁺* and *cry^b* flies is significant ($P < 0.0001$). Rhythmicity index values for *cry⁺* and *cry^b* were not different ($P > 0.1$). *luc* expression levels are different in *cry⁺* and *cry^b* antennae (see Supplementary Information).

‡ The difference between rhythmic *cry⁺* and *cry^b* flies is significant ($P < 0.002$). Rhythmicity index values for *cry⁺* vs. *cry^b* were not different ($P > 0.5$).

after entrainment by either photic or non-photic cues (Table 1) may be due to this being a non-null, missense mutant⁸. However, in whole-fly monitorings of *per-luc* and *tim-luc* rhythms, the mutation eliminates all such rhythmicity⁸. The low proportions of mutant rhythms in the current antennal monitorings are likely to have been revealed because different isolated tissues display separate phases of bioluminescence (Table 1 and Supplementary Information). Inasmuch as *cry^b* causes severe diminutions of rhythmicity in all tissues examined (Table 1 and Supplementary Information), the whole-animal result would be apparently aperiodic owing to temporal smearing of weak, unphased rhythms.

Previous studies showed that *cry^b* flies show locomotor activity rhythms in LD and DD, as well as robust PER cycling in the lateral-brain neurons that control these behavioural rhythms⁸. In contrast, the present results demonstrate abnormal antennal bioluminescence rhythms under a variety of conditions in *cry^b* flies and the loss of a physiological clock output in this tissue. These data indicate that CRY has a photoreceptor-independent function in *Drosophila* and further indicate that the role of CRY in circadian systems may be tissue specific. CRY could be a component of core oscillators throughout the periphery, or a component of a non-photic entrainment pathway such as thermoreception. However, there is no evidence of, or precedent for, CRY performing any thermoreceptive functions. We favour a role for CRY as a clock component of peripheral tissues, in line with the situation in mammals where CRY1 and CRY2 are essential components of the core oscillator in mouse⁶. Moreover, CRY can interact with TIM protein in a light-dependent manner¹⁸, which points to a possible mechanism by which CRY affects oscillator function in *Drosophila*. In this scenario, CRY could interact with TIM in peripheral tissues irrespective of the presence of light¹⁸, by analogy to the manner by which CRYs interact with other clock proteins in the central pacemaker of mouse^{6,19}. Note that a role for CRY in the core timekeeping mechanism does not prohibit this protein from functioning as a circadian photoreceptor in *Drosophila*, just as TIM's role in light-dependent entrainment (for example, refs 8, 18) does not preclude it from functioning as a core component of the oscillator. Thus, the versatility of CRY functioning in *Drosophila* is very similar to that in mouse, although the tissue situation is reversed: a function rather near the heart of circadian pacemaking in the periphery, and a photoreceptive function within the brain. □

Methods

Electroantennograms

We measured EAG responses to ethyl acetate under LD, DD, temperature cycles and constant temperature in females. The wild-type and *cry^b* strains used were as described⁸. EAG recordings made under LD and DD conditions were recorded and analysed as described¹³. For EAG studies using temperature entrainment, we reared flies in incubators in constant darkness under temperature cycles of 12 h 18°C:12 h 27°C. Transitions between high and low temperatures took <20 min. We made recordings under free-running conditions during the first day of constant high or constant low temperature. Results obtained on the second day of constant temperature were similar (data not shown).

Luciferase monitoring of clock gene expression

The transgenic *per-luc* (BG-*luc*) and *tim-luc* fly strains used for bioluminescence have been described^{8,16} (Table 1 and Supplementary Information). We used two additional *tim-luc* lines, which gave indistinguishable results. Data from the different transgenic strains (the separate BG-*luc* or *tim-luc* ones) and types (BG-*luc* or *tim-luc*) in a given *cry* genetic background passed homogeneity tests ($P \approx 0.2$ in all cases) and were pooled. For LD and DD experiments, dissections were carried out under ambient light. For constant-temperature experiments, dissections were performed under an attenuated light source. Aside from this brief (~1-min) illumination, the antennae in the temperature experiments were maintained in DD. Given the effect of light on *Drosophila* locomotor activity rhythms, this light exposure is unlikely to affect phase or period²⁰. Luminometric monitoring under LD and DD was done as described^{14,15}. Luminometric monitoring in temperature cycles (18°C:27°C) or constant temperature (27°C) was carried out in an incubator with programmable lights and temperature. We applied two temperature-entrainment protocols. Flies were reared (1) in LD cycles then transferred to DD for two temperature cycles; during the third cycle the antennae were dissected and bioluminescence was monitored for two additional cycles before the samples were released into

constant temperature; or (2) in DD in temperature-cycles; following dissection, samples were treated as in the first protocol.

Analysis of molecular timecourses

To evaluate molecular timecourses we applied analytical tools using Matlab, incorporating FORTRAN code written by us. The first 13–15 h of data were discarded owing to overly variable and high counts (compare with refs 14, 15, 21). We also discarded data for samples that did not have counts that started and remained above 100 c.p.s 72–96 h into the experiment. Control wells containing only luciferin and medium gave ≤ 80 c.p.s. About 50% of all samples met these criteria and were analysed as follows.

To assure that apparent fluctuations were not random, the record from each sample was shuffled with respect to time using the 'randperm' routine from Matlab. Each sample record was shuffled four times. The real and randomized timecourses were tested for significant periodicity by autocorrelation²². Although a 95% confidence interval was computed, correlograms were interpreted in terms of their shape because the data sets were small (~180 data points per sample) and the correlogram is sensitive to nonlinear trends in the data (downward linear trends were removed). Thus, data showing a regular rise and fall in the correlogram plot were rhythmic^{15,22,23}. We assessed a rhythmicity index (with higher values implying stronger periodic fluctuations) for rhythmic samples. The third peak in the correlogram (Fig. 2, asterisk) gives the most reliable estimate of rhythmic strength^{24,25}. Residual trends in the data can distort the correlogram leading to negative index values even for data that are significantly rhythmic (see the right-most column of Table 1). We estimated the period of a given significant rhythm by maximum entropy spectral analysis (MESA) because it provides the most sensitivity and the highest resolution compared to other methods^{22,23,26}. A given MESA peak was used to estimate period. When multiple peaks were present the higher one was taken to estimate the primary periodicity. Peaks >40 h were considered spurious. When the correlogram indicated arrhythmicity the MESA peak was ignored. Timecourses were assigned to one of the following categories (Table 1): arrhythmic; arrhythmic and white-noise-like (Fig. 2e); rhythmic, with periods in the ranges 10–22, 22–26 or 26–40 h. Phase estimates for LD timecourses were computed by determining the best fitting line that crosses the falling phase of each luminescence peak at its halfway point. The program then found the time in hours corresponding to when the data crossed that line and computed an average phase (per cycle) for the sample. Average phase values for all samples of a given genotype were then calculated as means of means (Table 1 and Supplementary Information). Levels of *luc*-reported clock gene expressions were computed as the average value in c.p.s. per h per specimen. Means of these means were computed for specimens of a given genotype (Table 1 and Supplementary Information).

Received 12 January; accepted 9 March 2001.

- van der Horst, G. T. J. et al. Mammalian Cry1 and Cry2 are essential for maintenance of circadian rhythms. *Nature* **398**, 627–630 (1999).
- Vitaterna, M. H. et al. Differential regulation of mammalian Period genes and circadian rhythmicity by cryptochromes 1 and 2. *Proc. Natl Acad. Sci. USA* **96**, 12114–12119 (1999).
- Kume, K. et al. mCRY1 and mCRY2 are essential components of the negative limb of the circadian clock feedback loop. *Cell* **98**, 193–205 (1999).
- Griffin, E. A. J., Staknis, D. & Weitz, C. J. Light-independent role of CRY1 and CRY2 in the mammalian circadian clock. *Science* **286**, 768–771 (1999).
- Okamura, H. et al. Photic induction of *mPer1* and *mPer2* in *Cry*-deficient mice lacking a biological clock. *Science* **286**, 2531–2534 (1999).
- Shearman, L. P. et al. Interacting molecular loops in the mammalian circadian clock. *Science* **288**, 1013–1019 (2000).
- Selby, C. P., Thompson, C., Schmitz, T. M., Van Gelder, R. N. & Sancar, A. Functional redundancy of cryptochromes and classical photoreceptors for nonvisual ocular photoreception in mice. *Proc. Natl Acad. Sci. USA* **97**, 14697–14702 (2000).
- Stanewsky, R. et al. The *cry^b* mutation identifies cryptochrome as a circadian photoreceptor in *Drosophila*. *Cell* **95**, 681–692 (1998).
- Emery, P., So, W. V., Kaneko, M., Hall, J. C. & Rosbash, M. CRY, a *Drosophila* clock and light-regulated cryptochrome, is a major contributor to circadian rhythm resetting and photosensitivity. *Cell* **95**, 669–679 (1998).
- Ishikawa, T. et al. DCRY is a *Drosophila* photoreceptor protein implicated in light entrainment of circadian rhythm. *Genes Cells* **4**, 57–65 (1999).
- Emery, P. et al. *Drosophila* CRY is a deep brain circadian photoreceptor. *Neuron* **26**, 493–504 (2000).
- Emery, P., Stanewsky, R., Hall, J. C. & Rosbash, M. A unique circadian-rhythm photoreceptor. *Nature* **404**, 456–457 (2000).
- Krishnan, B., Dryer, S. E. & Hardin, P. E. Circadian rhythms in olfactory responses of *Drosophila melanogaster*. *Nature* **400**, 375–378 (1999).
- Plautz, J. D., Kaneko, M., Hall, J. C. & Kay, S. A. Independent photoreceptive circadian clocks throughout *Drosophila*. *Science* **278**, 1632–1635 (1997).
- Plautz, J. D. et al. Quantitative analysis of *Drosophila period* gene transcription in living animals. *J. Biol. Rhythms* **12**, 204–217 (1997).
- Stanewsky, R., Jamison, C. F., Plautz, J. D., Kay, S. A. & Hall, J. C. Multiple circadian-regulated elements contribute to cycling *period* gene expression in *Drosophila*. *EMBO J.* **16**, 5006–5018 (1997).
- Wheeler, D. A., Hamblen-Coyle, M. J., Dushay, M. S. & Hall, J. C. Behavior in light-dark cycles of *Drosophila* mutants that are arrhythmic, blind or both. *J. Biol. Rhythms* **8**, 67–94 (1993).
- Ceriani, M. F. et al. Light-dependent sequestration of TIMELESS by CRYPTOCHROME. *Science* **285**, 553–556 (1999).
- Field, M. D. et al. Analysis of clock proteins in mouse SCN demonstrates phylogenetic divergence of the circadian clockwork and resetting mechanisms. *Neuron* **25**, 437–447 (2000).
- Power, J., Ringo, J. & Dowse, H. The role of light in the initiation of circadian activity rhythms of adult *Drosophila melanogaster*. *J. Neurogenet.* **9**, 227–238 (1995).
- Brandes, C. et al. Novel features of *Drosophila per* transcription revealed by real-time luciferase reporting. *Neuron* **16**, 687–692 (1996).

22. Chatfield, C. *The Analysis of Time Series* (Chapman and Hall/CRC, London, 1999).
23. Dowse, H. B. & Ringo, J. M. The search for hidden periodicities in biological time series revisited. *J. Theor. Biol.* **139**, 487–515 (1989).
24. Dowse, H. *et al.* A congenital heart defect in *Drosophila* caused by an action-potential mutation. *J. Neurogenet.* **10**, 153–168 (1995).
25. Johnson, E., Ringo, J., Bray, N. & Dowse, H. Genetic and pharmacological identification of ion channels central to the cardiac pacemaker. *J. Neurogenet.* **12**, 1–24 (1998).
26. Dowse, H. B. & Ringo, J. M. Comparisons between “periodograms” and spectral analysis: apples are apples after all. *J. Theor. Biol.* **148**, 139–144 (1991).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank R. Stanewsky for comments on the manuscript and additional *tim-luc* lines. This work was supported by US NIH grants to J.C.H., P.E.H. and S.E.D.

Correspondence and requests for materials should be addressed to P.E.H. (e-mail: phardin@uh.edu).

Integrins mediate functional pre- and postsynaptic maturation at a hippocampal synapse

Pascale Chavis* & Gary Westbrook

Vollum Institute Oregon Health Sciences, University Portland, Oregon 97201, USA

Coordinated signalling between presynaptic terminals and their postsynaptic targets is essential for the development and function of central synapses. In addition to diffusible molecules¹, this bidirectional flow of information could involve direct interactions through cell-adhesion molecules^{2,3}. Here, we show that one class of cell-adhesion molecule, the integrins, are required for the functional maturation of hippocampal synapses *in vitro*. At immature synapses, a high probability of glutamate release (P_r) was correlated with the expression of postsynaptic NMDA (N-methyl-D-aspartate) receptors containing the NR2B subunit. The activity-dependent reduction in P_r and a switch in the subunit composition of synaptic NMDA receptors was prevented by chronic blockade with peptides containing the integrin-binding site Arg-Gly-Asp (RGD), or by a functional antibody against the $\beta 3$ integrin subunit. Active synapses, monitored by the uptake of antibodies against the intraluminal domain of synaptotagmin I, also had $\beta 3$ subunit immunoreactivity. Our results provide evidence that integrin-mediated signalling is essential for the orchestrated maturation of central excitatory synapses.

We measured the probability of glutamate release (P_r) at autaptic hippocampal synapses using progressive block of the NMDA-receptor-mediated excitatory postsynaptic currents (EPSCs) by the irreversible NMDA open-channel blocker MK-801⁴. At 6–8 DIV (days *in vitro*), the progressive block by MK-801 (5 μ M) displayed two components ($\tau_{\text{fast}} = 2.4 \pm 0.3$, $\tau_{\text{slow}} = 22.4 \pm 2.1$ stimuli; $n = 16/17$, Fig. 1a), consistent with release sites with different P_r ^{4,5}. On the basis of the tenfold difference in the time constants, these two components represent ‘high’ and ‘low’ P_r synapses, although the actual distribution of release probabilities may not be strictly bimodal⁶. The proportion of the fast component was $50.7 \pm 5.3\%$ ($n = 16/17$; Fig. 1d). This indicates that the two populations of release sites contributed equally to synaptic trans-

mission, and that low P_r synapses outnumbered high ones under basal conditions. To confirm that the two components reflect differences in P_r , we used a paired-pulse protocol. If high P_r terminals are blocked preferentially during the first few stimuli in MK-801, the paired-pulse ratio (P2/P1) for the NMDA-receptor-

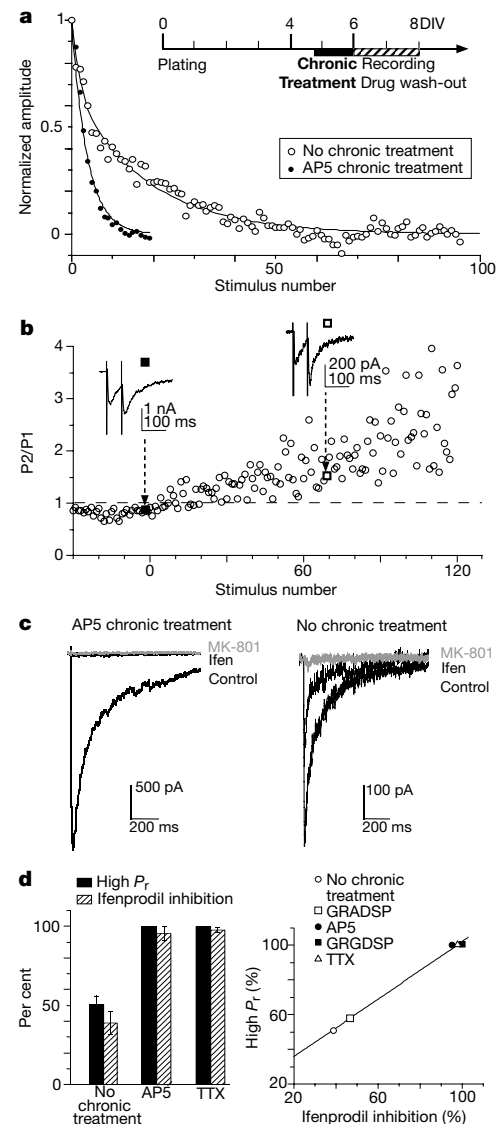


Figure 1 Presynaptic and postsynaptic properties are determined by synaptic activity.

a, Progressive block of the NMDA EPSC by a high concentration of MK-801 (5 μ M), as a function of stimulus number, fitted with two exponentials in an untreated hippocampal neuron ($\tau_{\text{fast}} = 2.1$ and $\tau_{\text{slow}} = 18.4$ stimuli; open circles). This is compared with a single fast exponential after chronic treatment with AP5 ($\tau_{\text{fast}} = 3.8$ stimuli; filled circles). **b**, Paired-pulse ratio of NMDA EPSCs gradually increased during progressive block by MK-801 in an untreated neuron. Before MK-801 and during the first stimulus in MK-801, the P2/P1 ratio was 0.82, but increased to 3.70 by stimulus 120 for this synapse. Insets, paired NMDA EPSCs at the beginning of MK-801 (filled squares) and at stimulus 70 (open squares). The interpulse interval was 60 ms. The progressive block rate by MK-801 also depends on the probability that NMDA channels will open (P_o) once bound by glutamate⁴. P_o did not change in our experiments (not shown). **c**, The NMDA EPSC was blocked completely by ifenprodil (ifen; 98.6%) after chronic treatment with AP5. After washout of ifenprodil, MK-801 progressively and totally blocked the synaptic response (grey trace). Ifenprodil was less effective in untreated cultures (28.6%, right trace). **d**, The percentage of high P_r synapses (solid bars) and sensitivity of the NMDA EPSC to ifenprodil (dashed bars) was increased markedly after block of synaptic activity. The percentage of high P_r synapses and the sensitivity to ifenprodil were tightly correlated ($r = 0.998$) after block of synaptic activity or inhibition of integrins (right; see also Fig. 2b).

* Present address: CNRS UPR 9023, 141 rue de la Cardonille, 34094 Montpellier Cedex 05, France.

mediated EPSC should increase with stimulus number. As seen in Fig. 1b, this was indeed the case. The P2/P1 was 0.81 ± 0.02 before MK-801 and 3.64 ± 1.63 after 120 stimuli in MK-801 ($n = 4$), confirming the validity of the MK-801 assay.

Chronic block of synaptic activity dramatically altered the P_r distribution. After chronic treatment with the NMDA receptor antagonist DL-2-amino-5-phosphonopentanoate (AP5), only the fast component was present ($\tau_{\text{fast}} = 3.48 \pm 0.21$, $n = 15/18$; Fig. 1a). Similar results were obtained after chronic blockade of neuronal activity with tetrodotoxin (TTX) ($\tau_{\text{fast}} = 3.96 \pm 0.34$, $n = 11/12$; Fig. 1d) but not with the AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid) receptor antagonist 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX) ($\tau_{\text{fast}} = 2.89 \pm 0.45$, $\tau_{\text{slow}} = 25.3 \pm 3.1$ stimuli, 64.9 $\pm$ 6.8% of high P_r synapses; $n = 7/7$, data not shown). The shift in the MK-801 progressive block to a single component could not be explained by changes in the properties of NMDA receptors, because the MK-801 block of whole-cell-evoked NMDA current had the same characteristics in control and in AP5-treated cells (Fig. 1). Thus, synapses that form in the absence of ongoing synaptic activity constitute a functionally uniform population with a high P_r .

A high P_r has been reported during early hippocampal synapse development⁷, but P_r decreases after nine DIV⁸. In 'mature' cultures (9–15 DIV), the progressive block of NMDA-mediated EPSC was fitted by a single slow component (20.54 ± 0.25 ; $n = 10$), consistent with a uniform population of release sites having a low P_r . This suggests that differences in P_r reflect a maturational change in the behaviour of individual synapses. We therefore considered whether activity blockade arrests such synapses at an immature stage. Early in development, NMDA receptors consist primarily of NR2B-containing receptors⁹, whereas NR2A subunits are expressed later. We used this developmental switch in NMDA receptor composition as a marker of postsynaptic maturation. After chronic treatment with AP5, the selective NR2B antagonist ifenprodil (3 μM)¹⁰ produced a nearly complete block of NMDA-mediated EPSCs ($95.5 \pm 4.5\%$, $n = 7$; Fig. 1c, d), but it blocked less than half of the EPSC at untreated synapses ($39.1 \pm 7.2\%$, $n = 10$). This pharmacological distinction was supported by the markedly slower deactivation of EPSCs after treatment with AP5 (not shown), typical of NR1/2B receptors¹¹. These results show that high P_r terminals are apposed to postsynaptic membrane primarily containing NR2B subunits, as shown by the tight correlation between ifenprodil block and the fraction of high P_r sites (Fig. 1d). After 9 DIV, low P_r terminals mediate NMDA EPSCs with a lower ifenprodil sensitivity¹². Consistent with this, we found a developmental decrease in the ifenprodil sensitivity of steady-state currents evoked by NMDA ($82.0 \pm 1.9\%$, $n = 12$ at 6 DIV; $45.2 \pm 3.8\%$, $n = 7$ at 9 DIV; and $38.1 \pm 3.3\%$, $n = 8$ at 12 DIV (data not shown)).

The synaptic phenotype observed after block of activity might indicate that high P_r synapses with NR2B-containing receptors represent nascent synapses. If so, in untreated cells (after block of synapses with ifenprodil) low P_r terminals should dominate the remaining NMDA EPSC. This proved to be the case. In the presence of ifenprodil, the progressive block in MK-801 largely consisted of the slow component ($\tau_{\text{slow}} = 18.8 \pm 1.0$ stimuli, $n = 10$; Fig. 2a). Because MK-801 block is irreversible, washout of ifenprodil and subsequent block by MK-801 revealed only a fast component ($\tau_{\text{fast}} = 3.24 \pm 0.11$; $n = 6$), consistent with a pure population of ifenprodil-sensitive high P_r sites. This strong correlation suggests that, early in development, synapses display coordinated pre- and postsynaptic properties.

We considered several mechanisms that could link directly the activity-dependent shift in pre- and postsynaptic properties. The importance of cell-adhesion molecules (CAMs) in neural development and synaptic plasticity is well recognized^{13,14}. The integrin family of CAMs can mediate heterophilic cell–extracellular matrix or cell–cell interactions¹⁵. Integrin interactions can be disrupted in

some cases with peptides containing the RGD sequence found in the binding site of many integrin ligands¹⁶. Chronic treatment with GRGDSP mimicked the effect of AP5 treatment (Fig. 2b). The progressive block by MK-801 was fitted with a single fast component

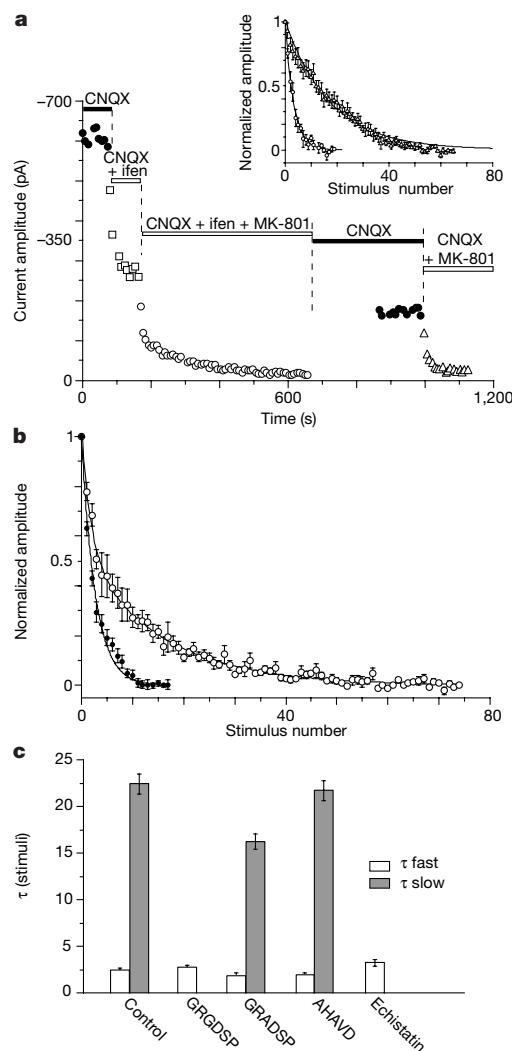


Figure 2 Block of integrins prevents synaptic maturation. **a**, Pure NR2B-containing synapses have a high P_r . In untreated cultures, ifenprodil (ifen, 3 μM) produced a 43.3% decrease in the NMDA EPSC measured in the presence of CNQX (open squares). The progressive block rate by MK-801 in the presence of ifenprodil (open circles) was best fitted by a single slow exponential ($\tau_{\text{slow}} = 15.1$ stimuli). The rapid block on the first stimulus in MK-801 probably represents a small population of partially ifenprodil-sensitive NR2A/2B receptors¹². After washout of ifenprodil (closed circles), the progressive block by MK-801 (open triangles) was fitted by a fast exponential ($\tau_{\text{fast}} = 2.1$ stimuli), consistent with unmasking of high P_r synapses. Inset, ensemble average $\pm$ standard error of the progressive block by MK-801 in the presence (τ_{slow} , open circle) and the absence (τ_{fast} , open triangle) of ifenprodil. **b**, Peptides interfering with adhesion of integrins mimic block of synaptic activity. Ensemble average $\pm$ standard error of the progressive block of the NMDA EPSC for synapses chronically treated with the integrin inhibitory peptide GRGDSP (50–200 μM ; $n = 14$; filled circles) or an inactive peptide (GRADSP 50–200 μM ; $n = 6$; open circles) are shown. A lower concentration of GRGDSP (5 μM) had no consistent effect. **c**, The shift in P_r was integrin specific. GRGDSP and the $\beta 1/\beta 3$ -specific integrin inhibitor (echistatin) eliminated the slow component of progressive MK-801 block, whereas GRADSP and the cadherin inhibitor peptide (AHAVD) had no effect. Control, $\tau_{\text{fast}} = 2.4 \pm 0.2$ and $\tau_{\text{slow}} = 22.4 \pm 2.1$ stimuli ($n = 16/17$); GRGDSP, $\tau_{\text{fast}} = 2.8 \pm 0.2$ stimuli ($n = 14/15$); GRADSP, $\tau_{\text{fast}} = 1.8 \pm 0.3$ and $\tau_{\text{slow}} = 16.3 \pm 1.2$ stimuli ($n = 7/7$); AHAVD, $\tau_{\text{fast}} = 1.9 \pm 0.2$ and $\tau_{\text{slow}} = 21.7 \pm 2.8$ stimuli ($n = 6/7$); echistatin, $\tau_{\text{fast}} = 3.2 \pm 0.3$ stimuli ($n = 6/7$).

($\tau_{\text{fast}} = 2.8 \pm 0.2$ stimuli, $n = 14/15$; Fig. 2c) and the NMDA EPSC was completely blocked by ifenprodil ($98.1 \pm 3.4\%$, Fig. 1d, right panel). Thus, chronic treatment with GRGDSP revealed a pure population of synapses with a high P_r and NR2B-containing receptors. A single amino-acid substitution in the canonical sequence (GRADSP) completely abolished peptide activity (Fig. 2b). After chronic treatment with GRADSP, the progressive block by MK-801, the percentage of high P_r sites ($58.2 \pm 5.1\%$) and the ifenprodil sensitivity of the NMDA EPSC ($45.9 \pm 2.1\%$; $n = 7/7$) were not significantly different from untreated cells (Figs 1d and 2c). Homophilic interactions of the classical cadherins have also been implicated in synaptic plasticity¹⁷. However, chronic treatment from DIV 0 or DIV 4.5 with a peptide containing the HAV sequence (AHAVD), which is known to interfere with their dimerization, had no effect (Fig. 2c). Thus, an RGD-sensitive integrin seems essential for activity-dependent synaptic maturation.

Integrins mediate specific physiological actions by restricted association of 16α - and 8β -subunits, resulting in 22 dimers with differing ligand specificities¹⁵. Several α and β subunits expressed in brain could form an RGD-sensitive integrin^{16,18}. The disintegrin echistatin, a peptide from viper venom containing an RGD sequence that inhibits $\beta 3$ - and $\beta 1$ -containing integrins¹⁹, was as effective as the active RGD peptide (Fig. 2c). To discriminate between $\beta 1$ and $\beta 3$, we used function-blocking antibodies. After chronic treatment with anti- $\beta 3$ antibody, the NMDA EPSC was entirely blocked by ifenprodil ($97.3 \pm 3.3\%$, $n = 6$; Fig. 3a), consistent with 'pure' NR2B-containing receptors. After removal of ifenprodil, the progressive block by MK-801 was fitted by one

Table 1 Transduction pathway linked to integrins

Drug	τ_{fast}	τ_{slow}	High P_r sites (%)	n
Genistein (50 μM)	3.57 ± 0.53	—	100	7/7
Genistein acute (50 μM)	1.96 ± 0.50	15.62 ± 4.24	50.2 ± 11.8	4/4
Daidzein (50 μM)	4.18 ± 1.72	22.71 ± 0.17	37.4 ± 9.0	5/5
AG 1478 (300 nM)	2.45 ± 0.6	20.77 ± 4.2	42.8 ± 8.4	7/7
PP2 (1–10 μM)	2.42 ± 0.32	24.10 ± 2.17	52.0 ± 5.5	11/11
K252a (17–50 nM)	3.48 ± 0.42	—	100	5/6
K252b (50–1,000 nM)	2.34 ± 0.89	25.27 ± 3.14	45.50 ± 18.12	4/4

The distribution of high and low P_r sites was assayed after chronic treatment (DIV 4.5 to DIV 6–8) with reagents affecting intracellular signalling cascades. Time constants of the progressive block by MK-801 corresponding to the high and low P_r sites are listed, as well as the fraction of high P_r (s.e.m.). Chronic, but not acute, inhibition of tyrosine kinases with genistein had the same effect as block of integrins. Daidzein, the inactive analogue of genistein, had no effect, as did inhibition of the EGF receptor with 10 μM PP2 or tyrphostin AG1478. Because BDNF enhances formation of new synapses through TrkB, a tyrosine kinase receptor, we examined the effect of K252a, a TrkB signalling blocker. We used K252b as a negative control. The MAP kinase inhibitor PD98059 (10 μM) reduced the P_r distribution to a single component, but the P_r was lower than that after tyrosine kinase inhibitors (data not shown).

exponential ($\tau_{\text{fast}} = 4.43 \pm 0.39$ stimuli, $n = 10/11$; Fig. 3a and inset). By contrast, chronic treatment with anti- $\beta 1$ antibody did not modify the phenotype of synapses ($\tau_{\text{fast}} = 3.6 \pm 0.6$ and $\tau_{\text{slow}} = 17.1 \pm 1.3$ stimuli, $n = 11/12$; Fig. 3a inset; ifenprodil inhibition was $46.4 \pm 2.0\%$, $n = 6$).

As shown in Fig. 3b–e, the $\beta 3$ function-blocking antibody labelled putative synapses along the neurites of hippocampal neurons. The uptake of an antibody targeted against the luminal epitope of the vesicular protein synaptotagmin I (StgI)²⁰ was used to visualize active, single synaptic boutons (Fig. 3c). $\beta 3$ -immuno-

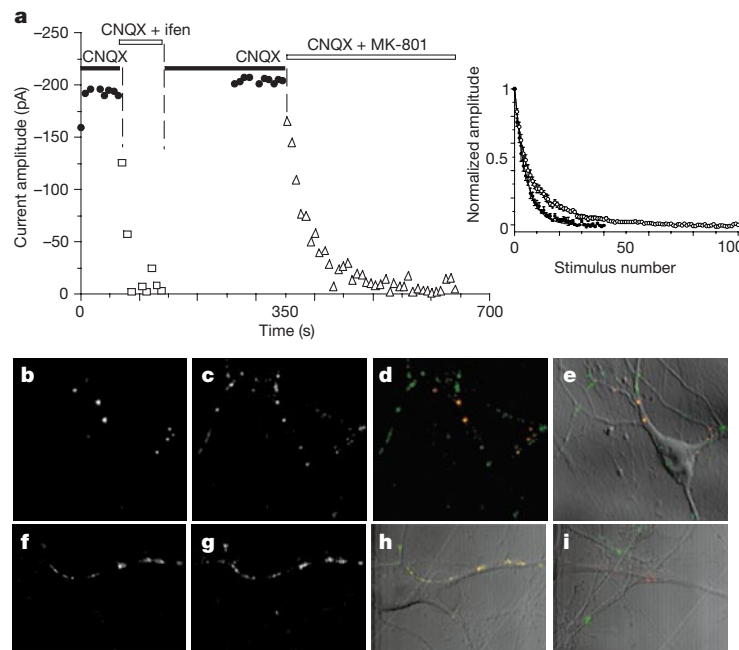


Figure 3 Role of $\beta 3$ -containing integrins in coordinated pre- and postsynaptic function. **a**, $\beta 3$ -integrin functional antibodies arrest synaptic maturation. Representative time course of MK-801 progressive block after chronic treatment with an anti- $\beta 3$ antibody. The EPSC was completely blocked by ifenprodil (ifen, 3 μM ; open squares). After wash-out of ifenprodil (filled circles), the time course of MK-801 block (open triangles) was fitted by one fast exponential ($\tau_{\text{fast}} = 4.68$ stimuli). Inset, normalized progressive block plots for 10 cells treated with anti- $\beta 3$ antibody (filled circle) and 11 cells chronically treated with anti- $\beta 1$ antibody (open circles). **b–e**, $\beta 3$ -containing integrins are localized at active synapses. Confocal microscopic images of immunostaining for the $\beta 3$ subunit (**b**) and StgI (**c**) on living hippocampal neurons. The dynamic uptake of anti-StgI antibodies allows visualization of individual active synapses. **d**, Superimposition of pseudocolour images **b**

(red) and **c** (green) indicates localization (yellow dots) of $\beta 3$ -IR at active synapses. Pseudocoloured images were superimposed on the Nomarski image to indicate the cellular distribution of both markers (**e**). Labelling was performed on high-density cultures to facilitate the identification of active synapses. Control experiments confirmed that activity blockade resulted in high ifenprodil sensitivity as expected for arrest of synaptic maturation (not shown). No immunoreactivity was observed with control isotype antibody. **f–i**, Localization of $\beta 3$ -containing integrins at specific synapses. **f–h**, Same as for **b–e**, but the labelling was performed on living hippocampal neurons chronically treated with AP5. **i**, Labelling with the NR2A NMDA-receptor subunit (green) and the $\beta 3$ subunit (red) are superimposed.

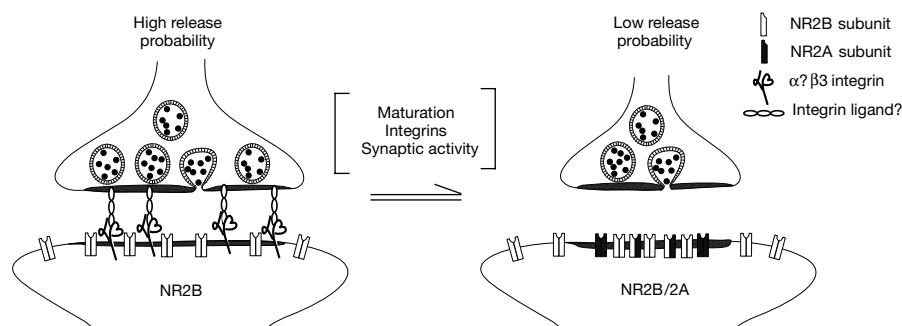


Figure 4 Working model for role of integrins in synaptic maturation. Integrins may be involved in coordinated changes in the pre- and postsynaptic properties of excitatory synapses. The integrin-dependent step must follow initial synapse formation and could involve either direct molecular interactions with a transsynaptic ligand or indirect interactions through a ligand in the extracellular matrix. The molecular mechanisms for

the maturation shift in P_r may involve a decrease in the size of the active zone, as has been measured at hippocampal synapses²¹. The switch in the postsynaptic NMDA receptor composition with incorporation of ifenprodil-insensitive synaptic receptors (i.e. NR2A-containing) would depend on intracellular transduction pathways mediated by integrin.

reactivity and StgI-immunoreactivity were colocalized at a subset of putative, active synapses on the soma and processes (Fig. 3e). These sites were preferentially located along proximal processes with 22.0% somatic, 46.6% between the soma and the first branch ($P = 0.0062$, Mann-Whitney U -test) and 31.3% after the first branching point. For 855 sites, $\beta 3$ -IR was detected only at active sites, although only 14.7% of the active sites showed $\beta 3$ -IR. The colocalization of $\beta 3$ - and StgI-IR was greatly increased after chronic treatment with AP5 (Fig. 3f–h, $51.6 \pm 1.9\%$ of colocalization, $n = 1,272$). A similar pattern was observed with another $\beta 3$ integrin antibody (not shown). No staining was observed on glial cultures. The localization of the $\beta 3$ subunit at a subset of synapses results from the heterogeneity of the synaptic sites. For example, inhibitory synapses were also present based on staining with GABA (γ -aminobutyric acid)-glutaraldehyde (not shown). No colocalization was seen between $\beta 3$ -IR sites and the NR2A subunit (Fig. 3i). These results suggest that the $\beta 3$ -containing integrins are preferentially expressed at immature sites, and that synaptic activity is not required for the synaptic localization of $\beta 3$ -IR. The anti- $\beta 1$ functional antibody also did not label synaptic sites, but $\beta 1$ -IR was colocalized with glial fibrillary acidic protein, suggesting a glial localization (not shown). We did not investigate the identity of the involved integrin subunit; however, $\beta 3$ forms heterodimers exclusively with α_{IIb} or α_v only the latter of which is expressed in brain.

In addition to their structural function in cell adhesion, integrins also activate intracellular signalling cascades. The most common integrin-mediated signalling pathway involves cytoplasmic tyrosine kinases (FAK or Src) and/or the mitogen-activated protein (MAP) kinase cascade²¹. We examined the potential function of tyrosine kinases in synaptic maturation, as summarized in Table 1. Chronic treatment (36 h) with genistein, a broad-spectrum tyrosine kinase inhibitor, mimicked the effect of the integrin inhibitors. However, we did not define the signalling cascade further as the Src inhibitor (PP2) was ineffective and another Src inhibitor, herbimycin, was lethal. Thus, integrin-mediated synaptic maturation requires tyrosine kinase activity, the identity of which remains unclear.

Our results are consistent with a model in which activity-dependent maturation of hippocampal excitatory synapses is dependent on integrin-mediated signalling. The structural and signalling features of CAMs such as the integrins are well suited to providing functional coupling of pre- and postsynaptic compartments. How might an integrin lead to a reduction in release probability? This may occur by altering the structure of the synapses, for example by reducing the overall area of synaptic contact, or by increasing the incidence of perforated synapses. As shown in the model in Fig. 4, we favour such a morphological explanation because immature hippocampal synapses *in vivo* and

in vitro are larger, and size has been suggested to explain the distribution of release probabilities^{22,23}. If the $\beta 3$ -containing integrin is postsynaptic, as shown, integrin-mediated coupling to tyrosine kinases could underlie the preferential localization of NR2A/2B receptors to mature synapses²⁴. Activity might also regulate the expression of $\beta 3$ -containing integrin such that the $\beta 3$ subunit is no longer expressed once the synapse is mature.

There are several candidate ligands for RGD-sensitive integrins^{2,25}, including L1, which is involved in neurite outgrowth and axon bundling, and also has a role in hippocampal synaptic plasticity²⁶. We propose a membrane-associated ligand as it provides the most direct link between observed pre- and postsynaptic maturation. A similar structural link, the neuroligin-neurexin complex, has been proposed as a trigger for synapse formation²⁷. In our experiments, integrins and synaptic activity were not required for synapse formation but rather for a subsequent maturational step that involved changes in the release probability and in the subunit composition at the postsynaptic density. Such a trans-synaptic junction provides a mechanism for continuous adaptation of the pre- and the postsynaptic properties to physiological events relevant for learning and memory. □

Methods

Cell culture and electrophysiology

Single, isolated hippocampal neurons were grown on 'microislands' as described⁴. We performed each set of experiments on three or more different cultures.

Autaptic currents were evoked by a 1-ms depolarizing voltage step from -70 mV to $+10$ mV, applied every 8–10 s. Pipette solution contained (in mM): KCl (140), EGTA (0.2), HEPES (10), glucose (10) and Mg_2ATP (2), pH 7.3 and osmolarity = 306–311 milliosmoles per litre. Control extracellular solution contained (in mM): NaCl (162), KCl (2.4), HEPES (10), glucose (10), glycine (0.02) and $CaCl_2$ (1.5), pH, 7.3 and osmolarity = 325 milliosmoles per litre. All solutions contained CNQX (10 μM), except when the AMPA-receptor-mediated EPSC was measured before and at the end of each experiment to monitor synaptic stability. We filtered the data at 2 kHz and digitized it at 4 kHz. Recordings were discarded if the access resistance was greater than 5 M Ω , as continuously monitored with 5 mV hyperpolarizing voltage pulses. Data were acquired with Clampex 7.0. and analysed with Clampfit 6.0.5 (Axon Instruments, Foster City, CA) and Kaleidagraph (Synergy Software). All data were plotted as a mean $\pm$ standard error and analysed with the Mann-Whitney U -test unless otherwise noted.

Chronic treatment of synapses

Reagents were added to the culture media 4.5 days after plating and were then removed by extensive washing just before electrophysiological recording (DIV 6–8) (Fig. 1a). Treatment from 0 DIV produced the same effect, presumably because synaptic activity was not prominent until DIV 6. Thus, chronic treatment did not prevent the formation of functional synapses. For longer treatments and to avoid possible peptide degradation, reagents were replenished every 2 days. Concentrations and sources of reagents were: AP5 100 μM (Tocris), CNQX 1–10 μM (Tocris), TTX 1 μM (Alomone Labs), GRGDSP 5–200 μM (Bachem) GRADSP 50 μM (Macromolecular Resources), Echistatin 0.1 μM (Sigma), AHAVD 200 μM –1 mM, hamster anti-rat $\beta 1$ monoclonal antibody 50–100 $\mu g\ ml^{-1}$ (Clone Ha2/5, Pharmingen), mouse anti-rat $\beta 3$ monoclonal antibody

50 $\mu\text{g ml}^{-1}$ (Clone F11, PharMingen), genistein 50 μM , daidzein 50 μM , AG1478 300 nM (Biomol), PP2 1–10 μM (Calbiochem), K252a 16–100 nM and K252b 50–1,000 nM (Biomol), herbimycin A (0.5–1 $\mu\text{g ml}^{-1}$ (Biomol).

Immunocytochemistry

To examine the expression of the $\beta 3$ integrin subunit, high density cultures were chronically treated with anti- $\beta 3$ antibody (25 $\mu\text{g ml}^{-1}$). After extensive washing with fresh medium, cells were incubated for 80 min at 37 °C with the rabbit anti-StgI antibody (1/100, Calbiochem) and a rhodamine-conjugated anti-mouse IgG (1/100, Chemicon). Cells were fixed in paraformaldehyde, permeabilized and incubated for 80 min with an Oregon Green-conjugated anti-rabbit IgG (Molecular Probes). Slides were viewed with a $\times 63$ objective lens on a laser confocal microscope (Odyssey XL, Noran Instruments).

Received 18 January; accepted 14 March 2001.

1. Fitzsimonds, R. M. & Poo, M. M. Retrograde signaling in the development and modification of synapses. *Physiol. Rev.* **78**, 143–170 (1998).
2. Kohmura, N. *et al.* Diversity revealed by a novel family of cadherins expressed in neurons at a synaptic complex. *Neuron* **20**, 1137–1151 (1998).
3. Song, J. Y., Ichtchenko, K., Sudhof, T. C. & Brose, N. Neuroligin 1 is a postsynaptic cell-adhesion molecule of excitatory synapses. *Proc. Natl Acad. Sci. USA* **96**, 1100–1105 (1999).
4. Rosenmund, C., Clements, J. D. & Westbrook, G. L. Nonuniform probability of glutamate release at a hippocampal synapse. *Science* **262**, 754–757 (1993).
5. Hessler, N. A., Shirke, A. M. & Malinow, R. The probability of transmitter release at a mammalian central synapse. *Nature* **366**, 569–572 (1993).
6. Murthy, V. N., Sejnowski, T. J. & Stevens, C. F. Heterogeneous release properties of visualized individual hippocampal synapses. *Neuron* **18**, 599–612 (1997).
7. Bolshakov, V. Y. & Siegelbaum, S. A. Regulation of hippocampal transmitter release during development and long-term potentiation. *Science* **269**, 1730–1734 (1995).
8. Chavis, P. & Westbrook, G. L. 'Pure' NR2B-containing synapses have a high release probability in cultured hippocampal neurons. *Soc. Neurosci. Abs.* (1998).
9. Williams, K., Russell, S. L., Shen, Y. M. & Molinoff, P. B. Developmental switch in the expression of NMDA receptors occurs *in vivo* and *in vitro*. *Neuron* **10**, 267–278 (1993).
10. Williams, K. Ifenprodil discriminates subtypes of the N-methyl-D-aspartate receptor: selectivity and mechanisms at recombinant heteromeric receptors. *Mol. Pharm.* **44**, 851–859 (1993).
11. Vicini, S. *et al.* Functional and pharmacological differences between recombinant N-methyl-D-aspartate receptors. *J. Neurophysiol.* **79**, 555–566 (1998).
12. Tovar, K. R. & Westbrook, G. L. The incorporation of NMDA receptors with a distinct subunit composition at nascent hippocampal synapses *in vitro*. *J. Neurosci.* **19**, 4180–4188 (1999).
13. Staubli, U., Chun, D. & Lynch, G. Time-dependent reversal of long-term potentiation by an integrin antagonist. *J. Neurosci.* **18**, 3460–3469 (1998).
14. Grotewiel, M. S., Beck, C. D., Wu, K. H., Zhu, X. R. & Davis, R. L. Integrin-mediated short-term memory in *Drosophila*. *Nature* **391**, 455–460 (1998).
15. Aplin, A. E., Howe, A., Alahari, S. K. & Juliano, R. L. Signal transduction and signal modulation by cell adhesion receptors: the role of integrins, cadherins, immunoglobulin-cell adhesion molecules, and selectins. *Pharmacol. Rev.* **50**, 197–263 (1998).
16. Ruoslahti, E. RGD and other recognition sequences for integrins. *Annu. Rev. Cell Dev. Biol.* **12**, 697–715 (1996).
17. Tang, L., Hung, C. P. & Schuman, E. M. A role for the cadherin family of cell adhesion molecules in hippocampal long-term potentiation. *Neuron* **20**, 1165–1175 (1998).
18. Pinkstaff, J. K., Dettmerich, J., Lynch, G. & Gall, C. Integrin subunit gene expression is regionally differentiated in adult brain. *J. Neurosci.* **19**, 1541–1556 (1999).
19. Pfaff, M., McLane, M. A., Bevilacqua, L., Niewiarowski, S. & Timpl, R. Comparison of disintegrins with limited variation in the RGD loop in their binding to purified integrins $\alpha 1 \text{Ib} \beta 3$, $\alpha 3 \text{Ib} \beta 3$, $\alpha 5 \text{Ib} \beta 3$ and $\alpha 5 \text{Ib} \beta 1$ and in cell adhesion inhibition. *Cell. Adhes. Commun.* **2**, 491–501 (1994).
20. Chavis, P., Mollard, P., Bockaert, J. & Manzoni, O. Visualization of cyclic AMP-regulated presynaptic activity at cerebellar granule cells. *Neuron* **20**, 773–781 (1998).
21. Schlaepfer, D. D. & Hunter, T. Integrin signalling and tyrosine phosphorylation: just the FAKs? *Trends Cell Biol.* **8**, 151–157 (1998).
22. Boyer, C., Schikorski, T. & Stevens, C. F. Comparison of hippocampal dendritic spines in culture and in brain. *J. Neurosci.* **18**, 5294–5300 (1998).
23. Schikorski, T. & Stevens, C. F. Quantitative ultrastructural analysis of hippocampal excitatory synapses. *J. Neurosci.* **17**, 5858–5867 (1997).
24. Lau, L. F. & Huganir, R. L. Differential tyrosine phosphorylation of N-methyl-D-aspartate receptor subunits. *J. Biol. Chem.* **270**, 20036–20041 (1995).
25. Blass, S., Kammerer, R. A. & Hall, H. Structural analysis of the sixth immunoglobulin-like domain of mouse neural cell adhesion molecule L1 and its interactions with $\alpha (v) \beta 3$, $\alpha (\text{IIb}) \beta 3$, and $\alpha 5 \beta 1$ integrins. *J. Neurochem.* **71**, 2615–2625 (1998).
26. Luthi, A., Laurent, J. P., Figurov, A., Muller, D. & Schachner, M. Hippocampal long-term potentiation and neural cell adhesion molecules L1 and NCAM. *Nature* **372**, 777–779 (1994).
27. Scheiffele, P., Fan, J., Choih, J., Fetter, R. & Serafini, T. Neuroligin expressed in nonneuronal cells triggers presynaptic development in contacting axons. *Cell* **101**, 657–669 (2000).

Acknowledgements

We thank G. Alonso, K. Tovar and A. Puel for discussions and O. Manzoni, C. Jahr and J. Bockaert for critical reading of the manuscript. We also thank A. Miller for technical support and M. Passama for the artwork. This work was supported by NIH (G.L.W.), the Human Frontiers Science Program, INSERM, and by FRM (P.C.).

Correspondence and request for materials should be addressed to P.C. (e-mail: chavis@ccpe.montp.inserm.fr).

LDL-receptor-related protein 6 is a receptor for Dickkopf proteins

Bingyu Mao*, Wei Wu*, Yan Li, Dana Hoppe, Peter Stannek, Andrei Glinka & Christof Niehrs

Division of Molecular Embryology, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

* These authors contributed equally to this work

Wnt glycoproteins have been implicated in diverse processes during embryonic patterning in metazoa. They signal through frizzled-type seven-transmembrane-domain receptors to stabilize β -catenin^{1,2}. Wnt signalling is antagonized by the extracellular Wnt inhibitor *dickkopf1* (*dkk1*), which is a member of a multigene family³. *dkk1* was initially identified as a head inducer in *Xenopus* embryos but the mechanism by which it blocks Wnt signalling is unknown. LDL-receptor-related protein 6 (LRP6) is required during Wnt/ β -catenin signalling in *Drosophila*, *Xenopus* and mouse, possibly acting as a co-receptor for Wnt^{4–6}. Here we show that LRP6 (ref. 7) is a specific, high-affinity receptor for Dkk1 and Dkk2. Dkk1 blocks LRP6-mediated Wnt/ β -catenin signalling by interacting with domains that are distinct from those required for Wnt/Frizzled interaction. *dkk1* and LRP6 interact antagonistically during embryonic head induction in *Xenopus* where LRP6 promotes the posteriorizing role of Wnt/ β -catenin signalling. Thus, DKKs inhibit Wnt co-receptor function, exemplifying the modulation of LRP signalling by antagonists.

We tested whether Dkk1 blocks Wnt signalling by binding to either Wnt proteins or Frizzled (Fz) receptors, but under various conditions we could detect no specific interactions. The ability of myc-tagged Xwnt8 protein⁸ to bind to Fz8 (ref. 8) in transfected 293T cells (Fig. 1a) was unaffected by Dkk1 protein (Fig. 1b) under conditions that completely blocked Xwnt8 signal transduction in a reporter assay (data not shown). In contrast, CRD8–Fc protein (the ligand-binding domain of *fz8* and a Wnt inhibitor that directly binds to Wnts and interferes with ligand–receptor interaction⁸) abolished binding of myc–wnt8 to *fz8*-transfected cells (Fig. 1c).

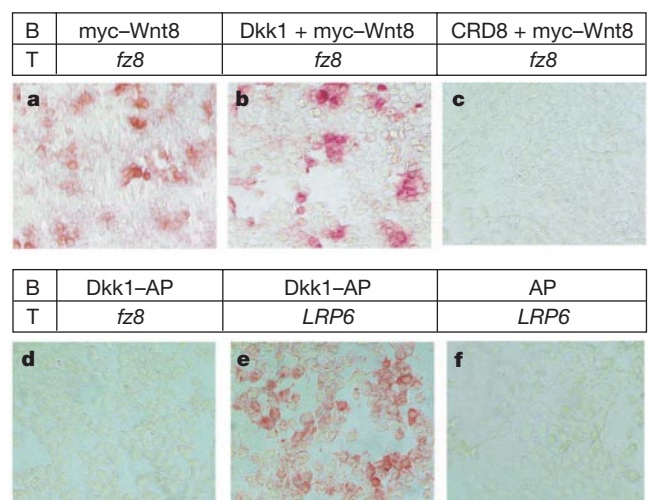


Figure 1 Dkk1 binds to LRP6-transfected cells. **a–c**, Wnt–Fz interaction is unaffected by Dkk1. 293T cells were transfected (T) with *mFz8*, and myc–Wnt8 protein with or without Dkk1 or mFz8CRD (CRD8) was applied for binding (B) as indicated. Cell-bound myc–Wnt8 was fixed and detected immunohistochemically. **d–f**, Dkk1–AP binds to LRP6 but not *mFz8*-transfected cells. 293T cells were transfected with *mFz8* or LRP6 and binding was done with Dkk1–AP or AP (alkaline phosphatase) protein; cells were stained for bound AP activity.

This indicated that Dkk1 may not interact directly with either Wnt or Fz proteins and raised the possibility that Dkk1 function requires another extracellular protein involved in Wnt signalling.

Human LRP6 is a putative co-receptor for Wnt signalling^{4–6}. We found that alkaline phosphatase (AP)-tagged Dkk1 (Dkk1-AP) binds effectively to 293T cells transfected with LRP6 (Fig. 1e). This interaction is specific, as Dkk1-AP does not bind to *fz8*-transfected cells (Fig. 1d) and AP does not bind to LRP6-transfected cells (Fig. 1f). Furthermore, both Dkk1 and Dkk2 (ref. 9) bind with high affinity (10^{-10} M) to LRP6-transfected cells, unlike the related Dkk3 protein⁹ (Fig. 2a,b). Consistent with this, binding of Dkk1-AP to LRP6-transfected cells was specifically inhibited by unlabelled Dkk1 and Dkk2, but not by Dkk3 (Fig. 2c).

To show that the interaction between Dkk1 with LRP6 is direct, we carried out a co-immunoprecipitation of Dkk1-AP with LRP6N-Fc⁵, a secreted protein consisting of the extracellular domain of LRP6 fused to an immunoglobulin Fc-tag. Dkk1-AP and Dkk2-AP were specifically precipitated by LRP6N-Fc, unlike Dkk3-AP (Fig. 2d). Control proteins included *Xenopus* Noggin-Fc¹⁰, which bound to neither Dkk1-AP nor Dkk2-AP, and cerberus-AP¹¹, which did not bind to LRP6. No interaction was detected between any Dkk-AP proteins and LDLR, VLDLR, ApoER2, LRP or LRP5 (not shown). We conclude that Dkk1 and Dkk2 are specific, high-affinity ligands for LRP6.

Next we analysed the functional interaction between LRP6 and Dkks during Wnt signalling. As read-out we used human HEK293T cells transfected with the luciferase reporter construct TOPFLASH, carrying T-cell factor (TCF)-binding sites, which is directly activated by the TCF/ β -catenin complex¹². In co-transfected 293T cells,

LRP6 induces the TOPFLASH reporter (Fig. 3a, lane 2) and synergizes with both *mwnt1* and *fz8* to activate Wnt signalling (Fig. 3a, lane 7), as shown in *Xenopus* embryos⁵. Signalling induced by *mwnt1*-*fz8* was inhibited by *dkk1* (Fig. 3b, lane 5) and *dkk2* (lane 6) but not by *dkk3* (lane 7). In contrast, *dkk1* was a weaker inhibitor against LRP6-*mwnt1*-*fz8* signalling than against the signal generated by *mwnt1*-*fz8* alone (lane 9). Neither *dkk2* nor *dkk3* inhibited LRP6-*mwnt1*-*fz8* at the applied dose (Fig. 3b, lane 10 and 11). This indicates that in the presence of transfected LRP6, Dkk1 may become limiting during inhibition of Wnt/ β -catenin signalling, presumably by becoming titrated. This was supported by the observation that *dkk1* inhibits signalling by high doses of LRP6 alone only poorly, whereas a *mwnt1*-*fz8* signal was very effectively inhibited (Fig. 3c, compare lanes 3, 4 with 8, 9), probably because *dkk1* overrides the low LRP levels endogenously present in 293T cells. This contrasts with mouse *FRP2*, which encodes a secreted Wnt inhibitor⁸ that acts by direct binding of Wnt protein and robustly inhibits LRP6 signalling (Fig. 3c, compare lanes 5, 6 with 10, 11). We conclude that LRP6 can titrate Dkk1 during inhibition of Wnt signalling.

To test whether LRP6 also functionally interacts with *dkk1* in *Xenopus*, we carried out animal cap assays in embryos for the induction of the immediate early Wnt signalling target *siamois* (*sia*)^{13,14}. At both low and high doses, injected LRP6 messenger RNA induced *sia*, as shown previously⁵ (Fig. 3d, lanes 7, 11). At a low LRP6 dose, *dkk1* inhibited the induction (lane 8), but *dkk3* had no effect (lane 10). At a high LRP6 dose, *dkk1* failed to inhibit *sia* induction (compare lanes 11 and 12), indicating that LRP6 can titrate Dkk1 in *Xenopus* as it does in 293T cells. *dkk2* strongly

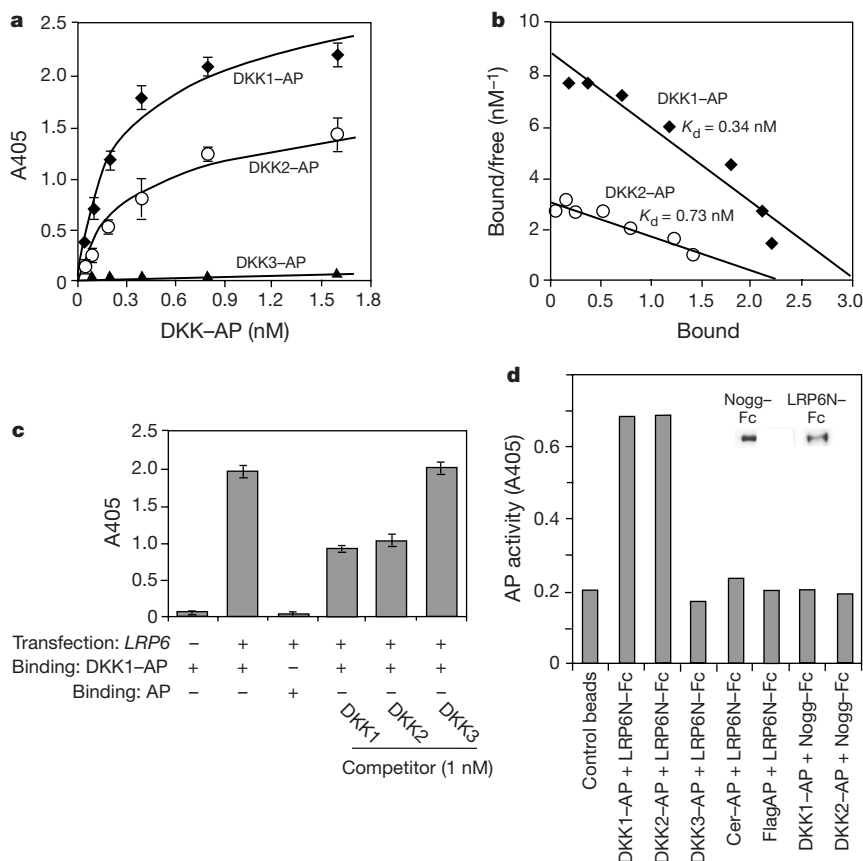


Figure 2 Dkk1 and Dkk2 are high-affinity ligands for LRP6. **a**, Binding of Dkk-AP fusion proteins to 293T cells transfected with LRP6. Dkk3-AP does not bind to LRP6-transfected cells. **b**, Scatchard analysis of the binding data shown in **a**. **c**, The binding of Dkk1-AP to LRP6-transfected cells can be specifically inhibited by unlabelled Dkk1 and Dkk2 but not Dkk3. **d**, Dkk1-AP and Dkk2-AP but not Dkk3-AP were specifically

precipitated by LRP6N-Fc. Indicated proteins were incubated *in vitro* and precipitated with protein G agarose. Co-immunoprecipitated AP activity was measured colorimetrically. The control proteins cerberus-AP (cer-AP) and Noggin-Fc (Noggin-Fc) do not co-immunoprecipitate with LRP6N-Fc or Dkk-AP, respectively. Inset, loading controls for LRP6N-Fc and Noggin-Fc.

synergized with *LRP6* in *sia* induction (lanes 9, 13), consistent with our previous finding that in early *Xenopus* embryos, unlike in 293T cells, *dkk1* and *dkk2* function antagonistically, which may result from *dkk2* repressing an inhibitory Wnt signalling pathway in *Xenopus*¹⁵.

The hallmark of *dkk1* action during early *Xenopus* embryogenesis is its ability to induce big heads when overexpressed on its own and to induce ectopic heads when expressed in conjunction with bone morphogenetic protein (BMP) inhibitors³. This is because Dkk1 functions in head induction by interfering with a posteriorizing Wnt signal during gastrulation¹⁶. If *dkk1* acts by inhibiting *LRP6* function, then a dominant-negative *LRP6*, *LRP6ΔC*, in which the intracellular domain of LRP6 is deleted and which blocks signalling of wild-type LRP6⁵, should mimic the effects of *dkk1*. Figure 4 (top) shows that injection of *LRP6ΔC* mRNA induces enlarged heads and big cement glands (78%, *n* = 67), as does *dkk1* (89%, *n* = 37). Furthermore, co-injection of *LRP6ΔC* with dominant-negative BMP receptor (tBR) mRNA induces ectopic heads (31%, *n* = 74), unlike tBR alone (0%, *n* = 62; Fig. 4, middle row). These results indicate that an endogenous LRP is normally involved in antagonizing *Xenopus* head formation, presumably by mediating signalling of a posteriorizing Wnt ligand. In support of this hypothesis, overexpression of wild-type *LRP6* in neuroectoderm by injection of animal blastomeres induces microcephalic embryos (59%, *n* = 51), thus mimicking the effect of expressing Wnts after mid-blastula transition¹⁷. Co-injection of *dkk1* mRNA with *LRP6* effectively rescues microcephaly (0%, *n* = 36) and thus inhibits LRP6 function during head development (Fig. 4, bottom row).

These results indicate that LRP6 or a related protein is normally involved in antagonizing *Xenopus* head induction, and that *dkk1* interferes with the function of *LRP6* in a fashion that is consistent

with Dkk1 protein acting as an inhibitory ligand of LRP6.

LRP6 is a multidomain protein, containing four extracellular amino-terminal epidermal growth factor (EGF)-like repeats, followed by LDL-receptor (LDLR) type A repeats that precede the transmembrane domain (Fig. 5a). To determine which of these domains interacts with Dkk1 we analysed various LRP6 deletion constructs for their ability to bind Dkk1-AP following transfection in 293T cells. We tested which of the deletions promotes Wnt signalling by itself as well as in synergy with co-transfected *mwnt1-fz8*. All deletions induced Wnt signalling on their own, with the exception of LRP6ΔC, which is dominant negative⁵ (Fig. 5c). The smallest deletion resulting in loss of Dkk1-AP binding was LRP6ΔE3–4, indicating that the domain encompassing EGF repeats 3 and 4 is required for interaction with Dkk1 (Fig. 5b). Conversely, the smallest deletion resulting in complete loss of ability to synergize with *mwnt1-fz8* was LRP6ΔE1–2 (Fig. 5c; note that the signal in the respective grey column is merely the sum of LRP6ΔE1–2 and *mwnt1-fz8* signals). This indicates that domains of LRP6 interacting with Dkk1 and with Wnt-Fz can be separated into E3–4 and E1–2, respectively. Interestingly, LRP6ΔE1–4 behaved like a constitutively active form, strongly activating the Wnt reporter even without co-transfected *mwnt1-fz8*. This indicates that the E1–4 domain may prevent LRP6 activation in the basal state and may work as an autoinhibitory domain. We also note that the intracellular domain of LRP6 has an important role during Wnt signalling. This is because LRP6ΔN, lacking essentially all extracellular domains, still weakly activated Wnt signalling, and LRP6ΔC, lacking the intracellular domain, behaved as a dominant negative as described previously⁵ (Fig. 5c and data not shown).

The conclusion that Wnt-Fz interaction is mediated by the E1–2 domain is supported by results obtained with a similar set of

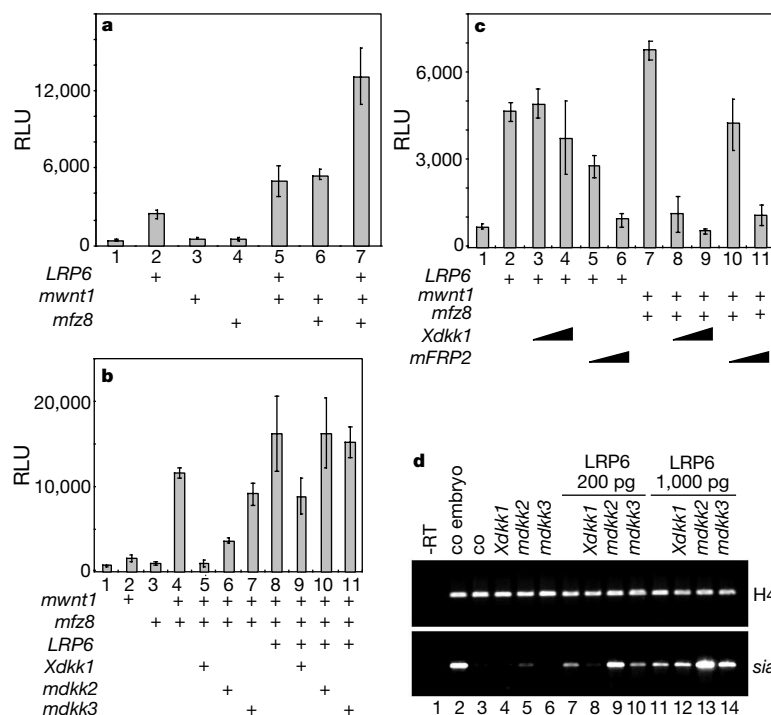


Figure 3 *LRP6* and *dkk1* functionally interact. **a**, *LRP6* synergizes with *wnt-fz* in a luciferase reporter assay in 293T cells. Note that sub-threshold doses of both *mwnt1* and *mfz8* were applied, which are not significantly active by themselves, to analyse the synergy with *LRP6*. RLU, relative luciferase units. **b**, Dkk1 inhibits effectively *wnt-fz* signalling but can only weakly inhibit *LRP6-mwnt1-mfz8* signalling. DNA concentrations used: *LRP6*, 3 ng per well; *Xdkk1* (*Xenopus dkk1*), *mdkk* (murine *dkks*) all at 25 ng per well. **c**, *dkk1* exerts a weaker inhibition against *LRP6* than against *mwnt1-mfz8* signalling

in reporter assay. DNA concentrations used: *Xdkk1*, 5 and 25 ng per well; *mFRP2*, 10 and 50 ng per well. **d**, *dkk1* can inhibit *siamois* induction by low but not high doses of *LRP6* in *Xenopus* embryos. Four-cell embryos were injected anally with the mRNAs indicated. Animal caps were explanted at stage 8 and *siamois* (*sia*) expression was analysed by RT-PCR at stage 10.5. mRNA concentrations used: *Xdkk1*, *mdkk2*, *mdkk3*, 250 pg per blastomere; *LRP6*, 200 or 1,000 pg per blastomere. H4, histone H4 for normalization.

deletion constructs created using the dominant-negatively acting LRP6 Δ C as backbone (Fig. 5d). We tested which extracellular domains in LRP6 Δ C are required for it to act as dominant negative and found that deletion of the E1–2 but not the E3–4 domain is sufficient to abolish its inhibitory action.

Wnt inhibitors of the sFRP^{18,8}, Cerberus¹⁹ and WIF²⁰ classes all function by directly binding to and sequestering Wnt proteins. In contrast, Dkks employ an indirect mode of action by binding to the

putative Wnt co-receptor LRP6. Lipoprotein receptors have traditionally been considered as merely mediating endocytosis of a variety of lipoproteins. However, the discovery of Reelin–ApoER2 signalling and Wnt–LRP6 signalling indicates that this large receptor family may have a widespread role in cell–cell communication²¹. Our finding that LRP6 is a Dkk receptor may therefore warrant screens for further antagonists of lipoprotein receptors and will guide searches for targets of other Dkks. □

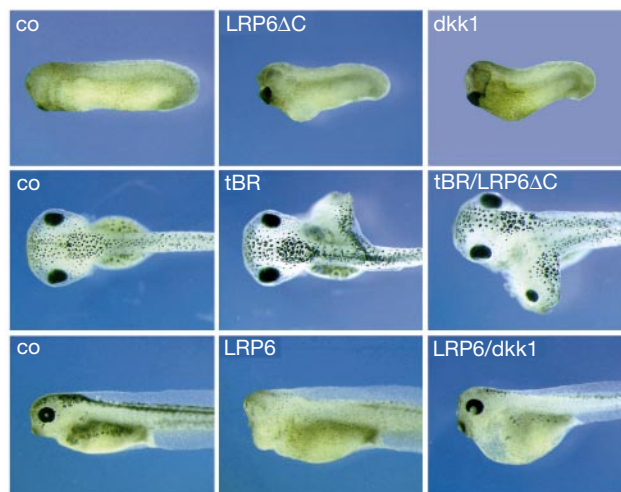


Figure 4 LRP6 and *dkk1* functionally interact in *Xenopus* head induction. Top, LRP6 Δ C anteriorizes *Xenopus* embryos like *Xdkk1*. co, control. mRNAs were radially injected at the four-cell stage in all blastomeres. Middle, LRP6 Δ C induces ectopic heads when co-injected with tBR. mRNAs were injected in one ventral blastomere at the 4–8-cell stage. Bottom, *dkk1* blocks posteriorization by LRP6. mRNAs were injected into four

animal blastomeres at the eight-cell stage embryos. Resulting tailbuds and tadpoles are shown in lateral view (top and bottom rows) and dorsal view (middle row). mRNA concentrations used (pg per blastomere): LRP6 Δ C, 500 (top); *Xdkk1*, 10 (top); tBR, 250; LRP6 Δ C, 250 (middle); LRP6, 1,000; *Xdkk1*, 250 (bottom).

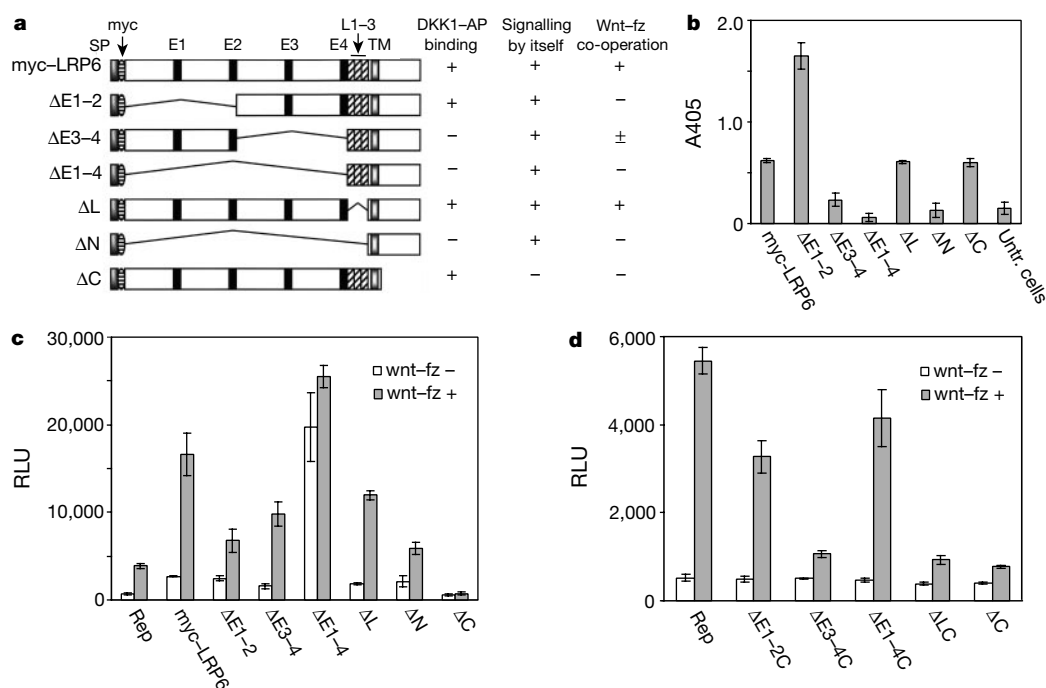


Figure 5 Distinct domains of LRP6 interact with Wnt–fz and Dkk1. **a**, Schematic drawing of LRP6 (myc tagged) and deletion constructs employed. Signal peptide (SP), EGF repeat domains (E1–E4), LDL repeats 1–3 (L1–3) and transmembrane domain (TM) preceding the cytoplasmic domain. Right, summary of properties of the deletion constructs in the assays shown in **b**, **c**. **b**, Binding of Dkk1–AP to 293T cells transfected with the deletion constructs shown in **a**. Untr. cells, untransfected cells. **c**, LRP6 deletion constructs in a luciferase reporter assay when co-transfected with (grey columns) or without (white

columns) *mwnt1*–*mfz8*. RLU, relative luciferase units. rep, reporter alone. **d**, Deletion of the E1–2 domain releases the dominant-negative effect of LRP6 Δ C. LRP6 Δ C blocks Wnt–fz signalling and this inhibition is released in Δ E1–2C and Δ E1–4C, but not other constructs. **c**, **d**, All LRP6 constructs were used at 3 ng per well; *mwnt1* was used at 8 ng per well. Expression of all deletion constructs was controlled by myc-epitope detection (not shown).

Methods

Embryos and explants

In vitro fertilization, embryo culture, staging, microinjection and culture of explants were carried out as described²².

Constructs

To generate AP (human placental alkaline phosphatase)-tagged *mdkk2*, *mdkk3* and *Xerberus*, the genes were cloned into pcDNA3-AP²³ in frame with carboxy-terminally added AP. pRKdkk1-AP was generated by cloning *Xdkk1* into a pRK5-based APTag2 vector²⁴ in frame with C-terminally added AP. pCSmyc-LRP6 was generated by fusing a cassette containing a heterologous signal peptide followed by a myc epitope (dip1-myc) to human LRP6⁵ and the construct was used as a template to construct LRP6 deletions, carried out by PCR. pCSmyc-LRP6ΔC was generated by deleting the C terminus C-terminally from residue 1,414 by PCR. ΔE1-2C and ΔLC were generated by PCR using ΔE1-2 and ΔL as template, ΔE1-4C and ΔE3-4C were generated by deleting sequences 3' of an internal *XhoI* site. pCSdkk1,2-flag were generated by fusing a flag epitope C-terminally and pCSflag-dkk3 was generated by inserting a flag epitope after the signal peptide of *mdkk3*. pcDNAMyc-AP was constructed by inserting the dip1-myc cassette in frame with AP.

Cell-surface binding

Protein conditioned media were produced by transient transfection of HEK293T cells with pRKdkk1-AP, pcDNAdkk2- and 3-AP, pCSdkk1,2-flag, pCSflag-dkk3, human pLRP6N-Fc⁵ and pcDNANoggin-Fc (*Xenopus* *noggin*; gift from S. Cohen), pRKmCRD8-Fc, pcDNAMyc-AP in serum-free medium (OPTIMEM I, Gibco). mycWnt8 protein was produced from S2 cells as described²⁵. Media were concentrated about 40-fold using Centricon Plus-20 filters (Millipore). The concentration of Dkk1-flag was determined by western blotting against an *in vitro*-translated Dkk1-flag standard which was quantified by ³⁵S-Met incorporation. The concentrations of Dkk1-AP, Dkk2-flag and flag-Dkk3 were determined by anti-Dkk1 or anti-flag western blotting against the quantified Dkk1. Dkk2,3-AP were quantified against Dkk1-AP by anti-AP western blotting. Protein production and integrity were controlled by western blotting. For cell-surface binding experiments, 293T cells were transfected using FuGENE 6 (Roche) for 48 h, incubated with fusion proteins for 1 h at room temperature, washed with PBS and stained with Fast Red (Roche). For binding curves and competition analysis, bound AP activity was measured as described²³.

Immunoprecipitation

AP fusion and Fc fusion proteins were produced as conditioned medium. For immunoprecipitation with protein-G agarose, equal amounts of AP- and Fc-tagged proteins were used as determined by AP activity or by western blotting. Conditions for incubation, immunoprecipitation and washing were as described¹⁹ except that 1 M NaCl was added during washing. Before colorimetric measurement of AP activity, immunoprecipitates were washed once with 5 mM Tris-HCl, pH 7.5, 1 mM MgCl₂ and 1 mM CaCl₂, and AP-substrate solution was added directly to the beads.

RT-PCR and luciferase assays

RT-PCR (PCR with reverse transcription) assays were carried out in the linear phase of amplification and with primers as described²⁶. Luciferase assays in HEK293T cells were done in 96-well plates at least in triplicate as described¹⁵. Plasmid concentrations used (ng per well): pTOPFLASH, 10; pTK-Renilla, 1; murine *Wnt1*, 5; *frizzled* 8, 1; *LRP6*, 10, unless indicated otherwise. DNA was adjusted to 100 ng DNA per well with pCS2+ vector DNA. Luciferase activity was normalized against *Renilla* activity.

Received 28 December 2000; accepted 5 April 2001.

- Wodarz, A. & Nusse, R. Mechanisms of Wnt signaling in development. *Annu. Rev. Cell. Dev. Biol.* **14**, 59–88 (1998).
- Miller, J. R., Hocking, A. M., Brown, J. D. & Moon, R. T. Mechanism and function of signal transduction by the Wnt/beta-catenin and Wnt/Ca²⁺ pathways. *Oncogene* **18**, 7860–7872 (1999).
- Glinka, A. *et al.* Dickkopf-1 is a member of a new family of secreted proteins and functions in head induction. *Nature* **391**, 357–362 (1998).
- Pinson, K. I., Brennan, J., Monksley, S., Avery, B. J. & Skarnes, W. C. An LDL-receptor-related protein mediates Wnt signalling in mice. *Nature* **407**, 535–538 (2000).
- Tamai, K. *et al.* LDL-receptor-related proteins in Wnt signal transduction. *Nature* **407**, 530–535 (2000).
- Wehrli, M. *et al.* arrow encodes an LDL-receptor-related protein essential for Wingless signalling. *Nature* **407**, 527–530 (2000).
- Brown, S. D. *et al.* Isolation and characterization of LRP6, a novel member of the low density lipoprotein receptor gene family. *Biochem. Biophys. Res. Commun.* **248**, 879–888 (1998).
- Rattner, A. *et al.* A family of secreted proteins contains homology to the cysteine-rich ligand binding domain of frizzled receptors. *Proc. Natl Acad. Sci. USA* **94**, 2859–2863 (1997).
- Monaghan, A. P. *et al.* Dickkopf genes are co-ordinately expressed in mesodermal lineages. *Mech. Dev.* **87**, 45–56 (1999).
- Smith, W. C. & Harland, R. M. Expression cloning of *noggin*, a new dorsalizing factor localized to the Spemann organizer in *Xenopus* embryos. *Cell* **70**, 829–840 (1992).
- Bouwmeester, T., Kim, S.-H., Sasai, Y., Lu, B. & De Robertis, E. M. Cerberus is a head-inducing secreted factor expressed in the anterior endoderm of Spemann's organizer. *Nature* **382**, 595–601 (1996).
- Korinek, V. *et al.* Constitutive transcriptional activation by a beta-catenin-Tcf complexes in APC/colon carcinoma. *Science* **275**, 1784–1787 (1997).

- Brannon, M. & Kimelman, D. Activation of Siamois by the Wnt pathway. *Dev. Biol.* **180**, 344–347 (1996).
- Carnac, G., Kodjachbajian, L., Gurdon, J. B. & Lemaire, P. The homeobox gene *siamois* is a target of the wnt dorsalization pathway and triggers organizer activity in the absence of mesoderm. *Development* **122**, 3055–3065 (1996).
- Wu, W., Glinka, A., Delius, H. & Niehrs, C. Mutual antagonism between dickkopf1 and-2 regulates Wnt/beta-catenin signalling. *Curr. Biol.* **10**, 1611–1614 (2000).
- Niehrs, C. Head in the Wnt—the molecular nature of Spemann's head organizer. *Trends Genet.* **15**, 314–319 (1999).
- Christian, J. L. & Moon, R. T. Interactions between Xwnt-8 and Spemann organizer signaling pathways generate dorsoventral pattern in the embryonic mesoderm of *Xenopus*. *Genes Dev.* **7**, 13–28 (1993).
- Leyns, L., Bouwmeester, T., Kim, S.-H., Piccolo, S. & De Robertis, E. M. Frzb-1 is a secreted antagonist of wnt-signals expressed in the Spemann organizer. *Cell* **88**, 747–756 (1997).
- Piccolo, S. *et al.* The head inducer Cerberus is a multifunctional antagonist of Nodal, BMP and Wnt signals. *Nature* **397**, 707–710 (1999).
- Hsieh, J. C. *et al.* A new secreted protein that binds to Wnt proteins and inhibits their activities. *Nature* **398**, 431–436 (1999).
- Willnow, T. E., Nykjaer, A. & Herz, J. Lipoprotein receptors: new roles for ancient proteins. *Nature Cell Biol.* **1**, E157–162 (1999).
- Gawantka, V., Delius, H., Hirschfeld, K., Blumenstock, C. & Niehrs, C. Antagonizing the Spemann organizer: role of the homeobox gene *Xvent-1*. *EMBO J.* **14**, 6268–6279 (1995).
- Cheng, H. J., Nakamoto, M., Bergemann, A. D. & Flanagan, J. G. Complementary gradients in expression and binding of ELF-1 and Mek4 in development of the topographic retinotectal projection map. *Cell* **82**, 371–381 (1995).
- Cheng, H. J. & Flanagan, J. G. Identification and cloning of ELF-1, a developmentally expressed ligand for the Mek4 and Sek receptor tyrosine kinases. *Cell* **79**, 157–168 (1994).
- Hsieh, J. C., Rattner, A., Smallwood, P. M. & Nathans, J. Biochemical characterization of Wnt-frizzled interactions using a soluble, biologically active vertebrate Wnt protein. *Proc. Natl Acad. Sci. USA* **96**, 3546–3551 (1999).
- Glinka, A., Wu, W., Onichtchouk, D., Blumenstock, C. & Niehrs, C. Head induction by simultaneous repression of Bmp and wnt signalling in *Xenopus*. *Nature* **389**, 517–519 (1997).

Acknowledgements

We thank X. He for reagents and advice; J. Nathans and P. Ramulu for the unpublished pRKdkk1-AP and other plasmids; and H. Clevers, S. Cohen, J. Flanagan, J. Herz, R. Moon, S. Sokol and N. Ueno for other materials. This work was supported in part by the Deutsche Forschungsgemeinschaft.

Correspondence and requests for materials should be addressed to C.N. (e-mail: niehrs@dkfz-heidelberg.de).

The status of Wnt signalling regulates neural and epidermal fates in the chick embryo

Sara I. Wilson*, Anna Rydström*, Tolleiv Trimborn†, Karl Willert†, Roel Nusse†, Thomas M. Jessell‡ & Thomas Edlund*

* Department of Microbiology, Umeå University, S-901 87 Umeå, Sweden
† Howard Hughes Medical Institute, Department of Developmental Biology, Stanford University, School of Medicine, Beckman Center, Stanford, California 94305-5323, USA
‡ Howard Hughes Medical Institute, Dept. of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, USA

The acquisition of neural fate by embryonic ectodermal cells is a fundamental step in the formation of the vertebrate nervous system. Neural induction seems to involve signalling by fibroblast growth factors (FGFs) and attenuation of the activity of bone morphogenetic protein (BMP)^{1–4}. But FGFs, either alone or in combination with BMP antagonists, are not sufficient to induce neural fate in prospective epidermal ectoderm of amniote embryos^{1,3,4}. These findings suggest that additional signals are involved in the specification of neural fate. Here we show that the state of Wnt signalling is a critical determinant of neural and epidermal fates in the chick embryo. Continual Wnt signalling blocks the response of epiblast cells to FGF signals, permitting the expression and signalling of BMP to direct an epidermal fate.

Conversely, a lack of exposure of epiblast cells to Wnt signals permits FGFs to induce a neural fate.

Nuclear localization of the Wnt transduction protein β -catenin marks cells exposed to active Wnt signals^{5–7}. In chick gastrula embryos, β -catenin is detected in the nuclei of lateral epiblast tissue that acquires epidermal fate, but is excluded from the nuclei of medial epiblast cells that generate neural tissue⁸, prompting us to consider whether the differential distribution of Wnt signalling regulates neural differentiation. We assessed the pattern of expression of two Wnt genes, *Wnt3A* and *Wnt8C*, before the onset of gastrulation, as neural fate is initiated^{3,4}. Medial (M; prospective neural) and lateral (L; prospective epidermal) epiblast tissues⁹ were isolated from stage X–XIII chick embryos⁹ and analysed for Wnt gene expression by a polymerase chain reaction with reverse transcription (RT–PCR) assay³. Stage XII–XIII (M) epiblast explants did not express *Wnt3A* or *Wnt8C*, whereas stage X–XIII (L) explants expressed both *Wnt3A* and *Wnt8C* (Fig. 1a). The gene encoding a Wnt8 receptor, *Frizzled8* (ref. 10), was expressed in both stage XII (M) and stage X (L) explants (Fig. 1a). Thus, Wnts that signal through β -catenin^{10,11} are expressed in lateral but not medial epiblast cells.

To address whether Wnt signalling regulates epidermal and neural fates, we examined the differentiation of cells in stage XII (M) epiblast explants in response to *Wnt3A*, *Wnt8* and *Wnt5A*. Expression of the transcription factors *Sox2*, *Sox3* and *Otx2* defines neural cells^{12,13}, and expression of *Msx1/2*, *GATA2* and *Dlx5* defines cells of epidermal character^{14–16}. Stage XII (M) explants grown in isolation generated *Sox2*⁺/*Sox3*⁺/*Otx2*⁺ cells, and few, if any, *Msx1/2*⁺/*GATA2*⁺ cells (Fig. 2a–f). In the presence of *Wnt3A* or *Wnt8* (ref. 10), many *Msx1/2*⁺/*GATA2*⁺ cells, but no *Sox2*⁺/*Sox3*⁺/*Otx2*⁺ cells were generated (Fig. 2g–k; data not shown). In contrast, stage XII (M) explants exposed to *Wnt5A* still generated *Sox2*⁺/*Sox3*⁺/*Otx2*⁺ cells, and no *Msx1/2*⁺/*GATA2*⁺ cells (data not shown). Wnts also blocked neural and induced an epidermal character in medial epiblast cells in intact chick embryos maintained in New culture¹⁷ (see Supplementary Information Fig. 1). Thus, Wnts that act through β -catenin signalling are expressed in the lateral epiblast, and they block neural and promote epidermal fate in medial epiblast cells.

We next addressed how Wnt signals repress neural fate. During normal development, FGF signalling is required for the progression of epiblast cells to a neural fate^{13,14}. Inhibition of FGF signalling in medial epiblast cells by the FGF receptor antagonist SU5402 (ref. 18) blocks neural and promotes epidermal fates³. SU5402 exposure also induces *Bmp4* and *Bmp7* expression, and suppresses *Fgf3* expression³. Moreover, BMPs block neural differentiation and repress *Fgf3* expression in medial epiblast cells (Fig. 1b). Thus, FGF signalling promotes neural differentiation in medial epiblast cells, partly by suppressing *Bmp* gene expression.

We therefore examined whether the action of Wnts in repressing neural fate is mediated through the attenuation of epiblast cell responses to FGF signalling and consequent changes in *Bmp* gene expression. Wnts blocked neural differentiation in medial epiblast explants exposed to added FGF protein. Stage XII (M) explants exposed jointly to FGF2 (30 nM) and *Wnt3A* generated *Msx1/2*⁺ cells but no *Sox2*⁺/*Sox3*⁺/*Otx2*⁺ cells (Fig. 2l–p). *Wnt3A* and *Wnt8* also mimicked the action of SU5402 in maintaining *Bmp4* gene expression in stage XII (M) explants (Fig. 1c; and data not shown)³. In addition, exposure of stage XII (M) explants to *Wnt3A* mimicked SU5402 in suppressing *Fgf3* expression (Fig. 1c). These findings support the idea that Wnts block the response of epiblast cells to FGF signalling.

If *Bmp* gene expression and BMP activity select epidermal and neural fate at a step downstream of FGF signalling, agents that block the activity of BMP proteins might reverse the effects of FGF signal blockade, and restore neural fate in medial epiblast explants. Inhibitors of BMP activity—the dominant-negative BMP receptor

IA–Fc and IB–Fc fusion proteins (dnBMPR) or Noggin—restore neural fate in stage XII (M) explants exposed to 2 μ M SU5402 (ref. 3). We therefore examined whether these inhibitors of BMP activity also restored neural fate in medial epiblast cells exposed to *Wnt3A*. The joint addition of Noggin (or dnBMPR) and *Wnt3A* restored the generation of *Sox2*⁺/*Sox3*⁺/*Otx2*⁺ cells in stage XII (M) explants, and no *Msx1/2*⁺/*GATA2*⁺ cells were detected (Fig. 2q–u; and data not shown). These findings argue that Wnt regulation of neural differentiation is mediated, in part, through the regulation of *Bmp* expression, and also provide further evidence for the similarity of the actions of Wnts and FGF receptor antagonists in regulating neural differentiation.

The restoration of neural fate in medial epiblast cells by BMP inhibitors was dependent, however, on the concentration of added FGF receptor antagonist or Wnt. Whereas Noggin or dnBMPRs restored neural fate in stage XII (M) explants exposed to 2 μ M SU5402 (ref. 3), they did not restore the generation of *Sox2*⁺/*Sox3*⁺/*Otx2*⁺ cells in explants exposed to 5 μ M SU5402 (Fig. 2v–x, a'–c'; and data not shown). In these conditions *Msx1/2*⁺/*GATA2*⁺ cells were still generated (Fig. 2y, z, d', e'). Increasing the concentration of Noggin tenfold in the presence of 5 μ M SU5402 did not restore neural cell fate (data not shown). Similarly, in the presence of *Wnt3A* at a concentration threefold greater than the threshold for blockade of neural induction, addition of Noggin or dnBMPR was not able to restore *Sox2*⁺/*Sox3*⁺/*Otx2*⁺ cells (Fig. 2f'–o'; and data not shown).

These results could be explained if FGF signalling promotes neural differentiation by activating two transduction pathways:

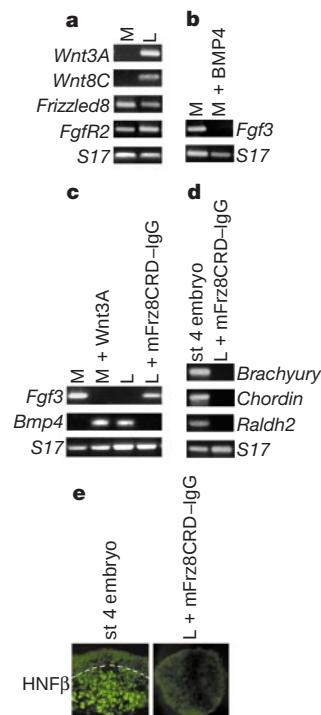


Figure 1 RT–PCR analysis of marker gene expression in medial and lateral epiblast explants. **a–d**, RT–PCR analysis of RNA isolated from stage X–XIII (EK X–XIII) explants. **a**, Expression of *Wnt3A*, *Wnt8C*, *Fgf receptor2* (*FgfR2*) and *Frizzled8* in stage XII–XIII (M) and stage X (L) explants before culture. *S17* expression is used as a control. **b**, *Fgf3* expression in stage XII–XIII (M) explants after 40 h of culture alone or with BMP4. **c**, *Fgf3* and *Bmp4* expression in stage XII–XIII (M) and stage X (L) explants grown alone or in the presence of *Wnt3A* or mFrz8CRD–IgG. **d**, Expression of mesodermal markers *Brachyury*, *Chordin* and *Raldh2* in Hamburger and Hamilton (HH) stage 4 (ref. 28) whole embryos but not in stage X (L) explants exposed to mFrz8CRD–IgG for 40 h. **e**, HNF3 β expression in HH stage 4 mesoderm but not in stage X (L) explants cultured with mFrz8CRD–IgG for 40 h.

one involving the repression of BMP expression; and one acting independently of the expression of BMPs. In this scheme, low concentrations of Wnts or SU5402 prevent neural differentiation primarily by blocking the FGF-mediated repression of BMP expression and activity: an action that can therefore be reversed by BMP inhibitors. But exposure of epiblast cells to high levels of Wnts or SU5402 would also block the second pathway, and thus these

inhibitory effects on neural differentiation are not reversed by BMP inhibitors.

This model argues that exposure to Wnt signals is a key constraint on the ability of lateral epiblast cells to progress to a neural fate. To test this idea, we cultured stage X (L) epiblast explants alone or in the presence of a soluble fragment of the mouse Frizzled protein (mFrz8CRD-IgG), an antagonist of Wnt signalling¹⁰. In stage X (L)

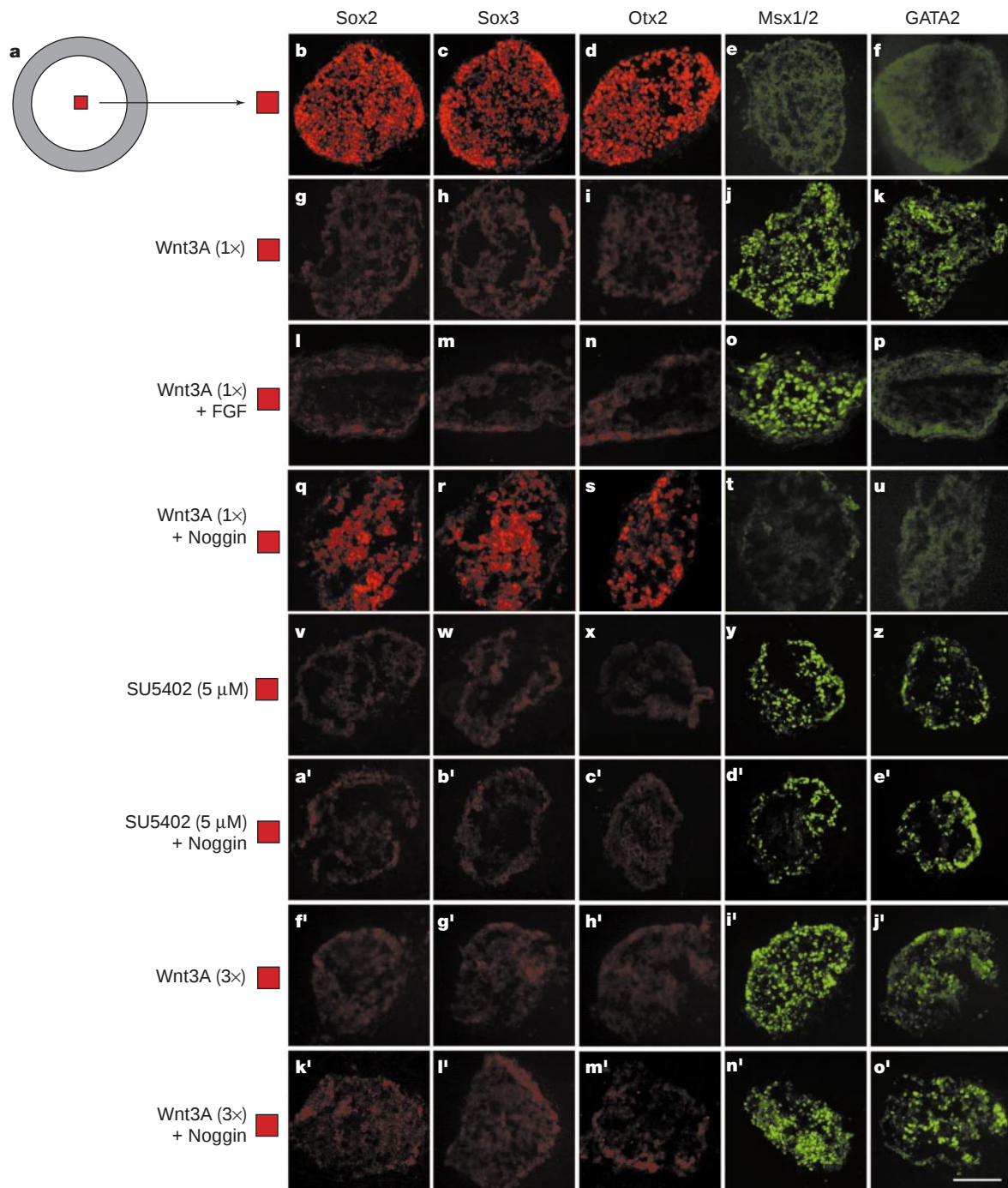


Figure 2 Regulation of neural and epidermal fate in medial epiblast explants. **a**, Representation of a stage XII embryo. Boxed area indicates the region of medial (M); red explants used for *in vitro* studies. **b–f**, Marker expression in medial explants cultured alone for 40 h. **g–k**, Marker expression in medial explants cultured in the presence of Wnt3A. **l–p**, Marker expression in medial explants cultured in the presence of Wnt3A and FGF2 (30 nM). No GATA2⁺ cells (**p**) are generated; this may result from FGF downregulation of GATA2 (ref. 29). **q–u**, Marker expression in medial explants cultured

with Wnt3A (1x) and Noggin (50 nM). **v–z**, Marker expression in medial explants cultured with SU5402 (5 μM). **a'–e'**, Marker expression in medial explants cultured with SU5402 (5 μM) and Noggin (50 nM). **f'–j'**, Marker expression in medial explants cultured in the presence of Wnt3A (3x). **k'–o'**, Marker expression in medial explants cultured with Wnt3A (3x) and Noggin (50 nM). See Supplementary Information for quantification of the data. Scale bar in **o'** (100 μm) applies to **b–o'**.

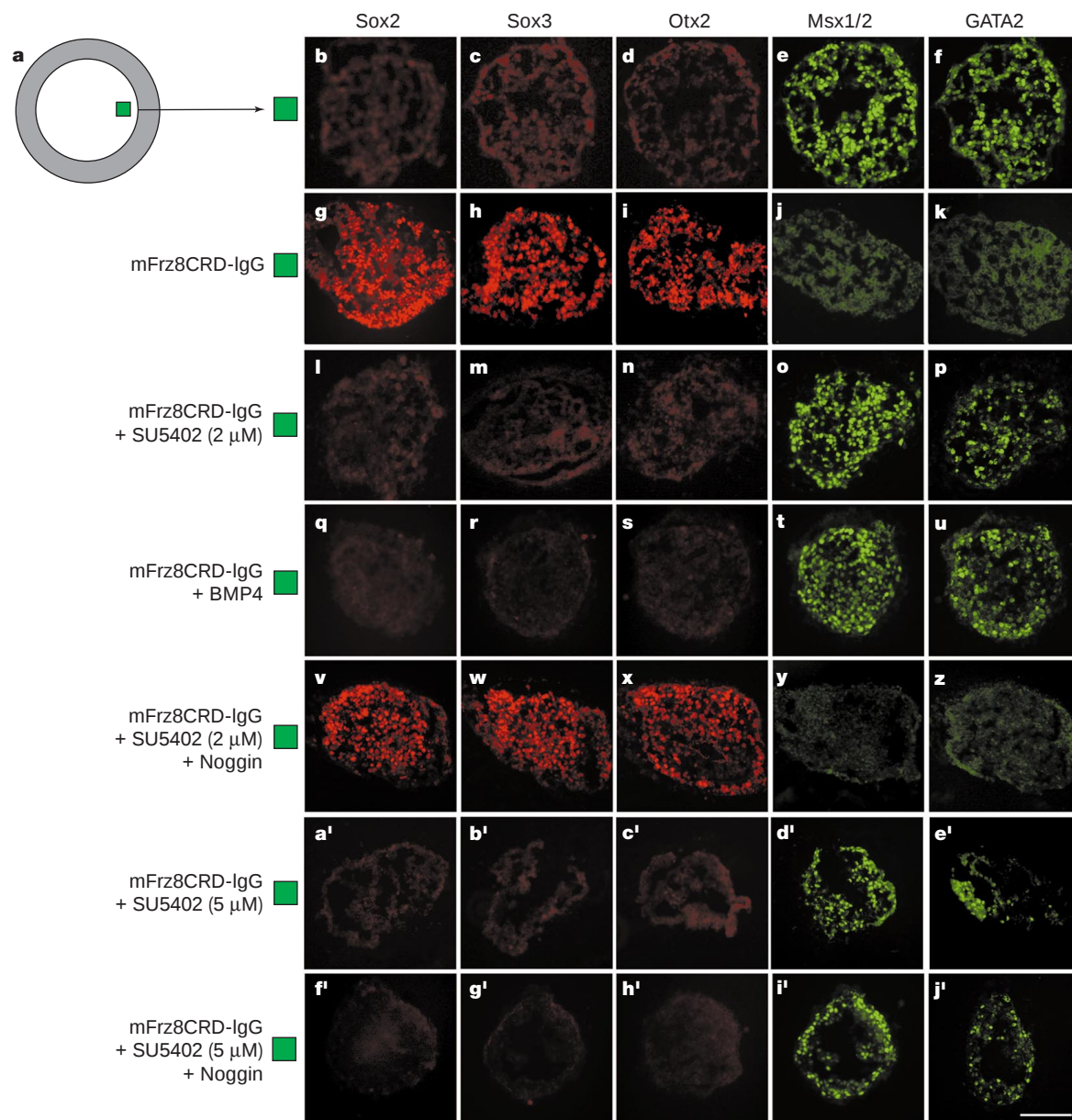


Figure 3 Regulation of neural and epidermal fate in lateral epiblast explants. **a**, Representation of a stage X embryo. Boxed area indicates region of lateral (L; green) explants used for *in vitro* studies. **b–f**, Marker expression in lateral explants cultured alone for 40 h. **g–k**, Marker expression in lateral explants cultured with mFrz8CRD–IgG. Explants isolated from the area opaca of stage X embryos grown in the presence of mFrz8CRD–IgG also generate Sox2⁺/Sox3⁺/Otx2⁺ cells, but not Msx1/2⁺/GATA2⁺ cells. **l–p**, Marker expression in lateral explants cultured with mFrz8CRD–IgG

and SU5402 (2 μM). **q–u**, Marker expression in lateral explants cultured with mFrz8CRD–IgG and BMP4. **v–z**, Marker expression in lateral explants cultured with mFrz8CRD–IgG, SU5402 (2 μM) and Noggin (50 nM). **a'–e'**, Marker expression in lateral explants cultured with mFrz8CRD–IgG and SU5402 (5 μM). **f'–j'**, Marker expression in lateral explants cultured with mFrz8CRD–IgG, SU5402 (5 μM) and Noggin (50 nM). Quantitation provided in Supplementary Information. Scale bar in **j'** (100 μm) applies to **b–j'**.

explants grown alone, Msx1/2⁺/GATA2⁺ cells but no Sox2⁺/Sox3⁺/Otx2⁺ cells were detected (Fig. 3a–f). In contrast, in the presence of mFrz8CRD–IgG, Sox2⁺/Sox3⁺/Otx2⁺ cells but no Msx1/2⁺/GATA2⁺ cells were detected (Fig. 3g–k). No mesodermal or endodermal cells were generated, as determined by the lack of expression of *Brachyury*, *Chordin*, *Raldh2* (ref. 3) or HNF3β¹⁹ (Fig. 1d, e). mFrz8CRD–IgG also induced neural and blocked epidermal character in prospective epidermal cells in chick embryos maintained in New culture (see Supplementary Information Fig. 1). Thus, inhibition of Wnt signalling in lateral epiblast cells results in neural differentiation.

We next examined whether the profile of gene expression in lateral epiblast cells that emerges after blockade of Wnt signalling conforms to the normal character of medial, prospective neural

epiblast cells. In stage X (L) explants grown in the presence of mFrz8CRD–IgG, the expression of *Bmp4* was suppressed and *Fgf3* was induced (Fig. 1c). To determine whether FGF signalling in lateral epiblast cells is necessary for the acquisition of neural character after elimination of Wnt signalling, we monitored neural differentiation in stage X (L) explants exposed to both mFrz8CRD–IgG and SU5402 (2 μM). In the presence of both agents, neural differentiation was blocked: Msx1/2⁺/GATA2⁺ cells but no Sox2⁺/Sox3⁺/Otx2⁺ cells were detected (Fig. 3l–p). Thus, FGF signalling is required for neural induction after inactivation of Wnt signalling. Blockade of Wnt signalling therefore initiates a program of neural differentiation in lateral epiblast cells that resembles the pathway normally followed by medial epiblast cells.

We also tested whether BMPs restore epidermal fate in lateral

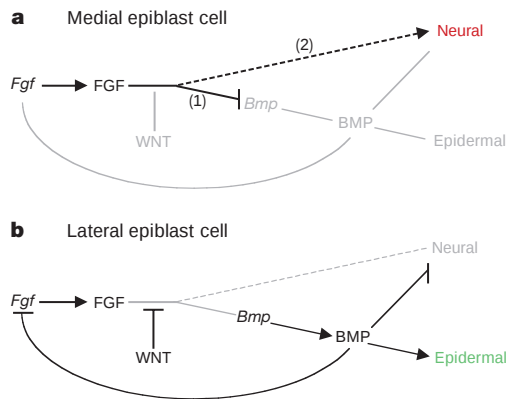


Figure 4 Proposed signalling pathway for neural induction in the chick embryo. **a**, Medial epiblast cells express FGFs but not Wnts. FGF signalling is shown to activate two distinct transduction pathways in epiblast cells: (1) repression of *Bmp* expression and prevention of BMP signalling (solid line from FGF), and (2) the promotion of a neural fate by a pathway independent of the repression of *Bmp* expression is required for the progression to a neural fate (dashed line from FGF). **b**, Lateral epiblast cells express FGFs and Wnts. Wnt signalling blocks the response of epiblast cells to FGFs, and thus *Bmp* genes are expressed. BMP signalling promotes epidermal fate and represses *Fgf* expression. Exposure to low concentrations of Wnts blocks the ability of FGFs to repress *Bmp* expression, but the independent pathway involved in promotion of neural fate is preserved. Thus, BMP antagonists can restore neural fate. High concentrations of Wnts block both FGF-dependent pathways, and thus BMP antagonists do not restore neural fate.

epiblast explants exposed to mFrz8CRD-IgG. In the presence of both BMP4 (2 nM) and mFrz8CRD-IgG, *Mx1/2*⁺/*GATA2*⁺ cells but no *Sox2*⁺/*Sox3*⁺/*Otx2*⁺ cells were detected (Fig. 3q–u). In lateral epiblast explants grown in the presence of mFrz8CRD-IgG and 2 μ M SU5402, addition of Noggin or dnBMPR led to the generation of *Sox2*⁺/*Sox3*⁺/*Otx2*⁺ cells but no *Mx1/2*⁺/*GATA2*⁺ cells (Fig. 3v–z). However, neither Noggin nor dnBMPR was able to induce neural properties in lateral explants exposed to mFrz8CRD-IgG and 5 μ M SU5402 (Fig. 3a'–j'). These results suggest why the addition of BMP antagonists alone does not induce neural character in stage X–XII (L) explants: presumably, cells in the lateral epiblast are normally exposed to a high level of Wnt activity that blocks the BMP-independent FGF transduction pathway required for neural differentiation. Consistent with this idea, a reduction in the level of Wnt signalling in stage X (L) explants, achieved by adding a concentration of mFrz8CRD-IgG threefold lower than the threshold for neural induction, permitted Noggin and dnBMPR to induce *Sox2*⁺/*Sox3*⁺/*Otx2*⁺ cells (see Supplementary Information Fig. 2).

Together, these findings support the idea that the state of Wnt signalling determines the selection of neural or epidermal cell fate in the chick epiblast. Our results can best be explained by a model in which the lack of exposure of medial epiblast cells to Wnts permits FGF signalling both to repress *Bmp* expression and to activate an independent pathway necessary for progression to neural fate (Fig. 4a). High-level Wnt signalling in lateral epiblast cells inhibits both FGF transduction pathways and permits *Bmp* expression and BMP signalling to direct cells to an epidermal fate (Fig. 4b). Resolving how spatial restrictions in *Wnt* expression are achieved, and how Wnt activity blocks the ability of ectodermal cells to respond to FGFs, may provide further insight into the sequential steps of neural induction. The *Sox1–3* proteins define early neural character^{12,20,21}, and seem to be involved in the control of neural differentiation^{21,22}. An interaction between *Sox3* and β -catenin has been reported to inhibit Wnt signalling²³, raising the possibility that extrinsic signals that select neural and epidermal fate converge at the level of regulation of *Sox* protein expression and activity. □

Methods

Isolation and growth of tissue explants

We isolated medial and lateral epiblast and area opaca explants from stage X–XIII (ref. 9) chick embryos and cultured them *in vitro*³.

Whole-embryo culture

Stage X–XI chick embryos were maintained in New culture¹⁷ for 22–24 h. Pellets of roughly 1,000 cells, transfected with a *lacZ* reporter, mouse *Wnt 3A* or *Frz8CRD*–IgG expression construct, were prepared as hanging-drop cultures 24 h after transfection. Cell aggregates were maintained in culture for 24 h, and then grafted into different regions of host embryos.

Wnt3A-, Wnt5A- and Wnt8-expressing cell lines

Mouse *Wnt3A* and mouse *Wnt5A* cDNAs were expressed under a PGK promoter and transfected into mouse lateral cells²⁴. The generation of the *Xenopus* Wnt8-expressing cells has been described¹⁰; these cells were kindly provided by J. Nathans.

Preparation of inducing factors

Soluble Wnt3A, Wnt5A and Wnt8 and control conditioned media were prepared as described^{10,24}. Soluble Frizzled 8 (mFrz8CRD–IgG) and control conditioned medium were prepared by transfecting HEK-293 cells with a *mfrz8CRD*–IgG (ref. 10) or *lacZ* reporter construct⁵. We used human BMP4 (R&D systems) and FGF2 (GibcoBRL) at 2 nM and 30 nM. Mouse BMP receptors IA/Fc and IB/Fc^{25,26} (R&D systems) were used at a final concentration of 0.5 μ M. Purified recombinant *Xenopus* Noggin, provided by R. Harland, was used at 50 nM. The FGF receptor inhibitor SU5402 (Calbiochem) was used at 2, 5 and 10 μ M.

Immunohistochemistry and *in situ* hybridization

We carried out immunohistochemistry and whole-mount *in situ* hybridization as described^{3,27}.

RT–PCR analysis

RT–PCR was performed on total RNA prepared from stage X–XIII medial and lateral explants ($n = 6–10$)³. Conditions and sequences of the oligonucleotide primers used in this study are given in Supplementary Information.

Received 22 January; accepted 19 March 2001.

- Harland, R. Neural induction. *Curr. Opin. Genet. Dev.* **10**, 357–362 (2000).
- Wilson, P. A. & Hemmati-Brivanlou, A. Induction of epidermis and inhibition of neural fate by Bmp-4. *Nature* **376**, 331–333 (1995).
- Wilson, S. I., Graziano, E., Harland, R., Jessell, T. M. & Edlund, T. An early requirement for FGF signalling in the acquisition of neural cell fate in the chick embryo. *Curr. Biol.* **10**, 421–429 (2000).
- Streit, A., Berliner, A. J., Papanayotou, C., Sirulnik, A. & Stern, C. D. Initiation of neural induction by FGF signalling before gastrulation. *Nature* **406**, 74–78 (2000).
- Noordermeer, J., Klingensmith, J., Perrimon, N. & Nusse, R. Dishevelled and armadillo act in the wingless signalling pathway in *Drosophila*. *Nature* **367**, 80–83 (1994).
- Kelly, G. M., Erezilymaz, D. F. & Moon, R. T. Induction of a secondary embryonic axis in zebrafish occurs following the overexpression of β -catenin. *Mech. Dev.* **53**, 261–273 (1995).
- Guger, K. A. & Gumbiner, B. M. β -Catenin has Wnt-like activity and mimics the Nieuwkoop signaling center in *Xenopus* dorsal–ventral patterning. *Dev. Biol.* **172**, 115–125 (1995).
- Roeser, T., Stein, S. & Kessel, M. Nuclear β -catenin and the development of bilateral symmetry in normal and LiCl-exposed chick embryos. *Development* **126**, 2955–2965 (1999).
- Eyal-Giladi, H. & Kochav, S. From cleavage to primitive streak formation: a complementary normal table and a new look at the first stages of the development of the chick. I. General morphology. *Dev. Biol.* **49**, 321–337 (1976).
- Hsieh, J. C., Rattner, A., Smallwood, P. M. & Nathans, J. Biochemical characterization of Wnt-frizzled interactions using a soluble, biologically active vertebrate Wnt protein. *Proc. Natl Acad. Sci. USA* **96**, 3546–3551 (1999).
- Shimizu, H. *et al.* Transformation by Wnt family proteins correlates with regulation of beta-catenin. *Cell Growth Differ.* **8**, 1349–1358 (1997).
- Rex, M. *et al.* Dynamic expression of chicken *Sox2* and *Sox3* genes in ectoderm induced to form neural tissue. *Dev. Dyn.* **209**, 323–332 (1997).
- Mallamaci, A., Di Blas, E., Briata, P., Boncinelli, E. & Corte, G. OTX2 homeoprotein in the developing central nervous system and migratory cells of the olfactory area. *Mech. Dev.* **58**, 165–178 (1996).
- Streit, A. & Stern, C. D. Establishment and maintenance of the border of the neural plate in the chick: involvement of FGF and BMP activity. *Mech. Dev.* **82**, 51–66 (1999).
- Sheng, G. & Stern, C. D. *Gata2* and *Gata3*: novel markers for early embryonic polarity and for non-neural ectoderm in the chick embryo. *Mech. Dev.* **87**, 213–216 (1999).
- Pera, E., Stein, S. & Kessel, M. Ectodermal patterning in the avian embryo: epidermis versus neural plate. *Development* **126**, 63–73 (1999).
- New, D. A. A new technique for the cultivation of the chick embryo *in vitro*. *J. Embryol. Exp. Morphol.* **3**, 320–331 (1955).
- Mohammadi, M. *et al.* Structures of the tyrosine kinase domain of fibroblast growth factor receptor in complex with inhibitors. *Science* **276**, 955–960 (1997).
- Dufort, D., Schwartz, L., Harpal, K. & Rossant, J. The transcription factor HNF3beta is required in visceral endoderm for normal primitive streak morphogenesis. *Development* **125**, 3015–3025 (1998).
- Wood, H. B. & Episkopou, V. Comparative expression of the mouse *Sox1*, *Sox2* and *Sox3* genes from pre-gastrulation to early somite stages. *Mech. Dev.* **86**, 197–201 (1999).
- Kishi, M. *et al.* Requirement of *Sox2*-mediated signaling for differentiation of early *Xenopus* neuroectoderm. *Development* **127**, 791–800 (2000).

22. Pevny, L. H., Sockanathan, S., Placzek, M. & Lovell-Badge, R. A role for SOX1 in neural determination. *Development* **125**, 1967–1978 (1998).
23. Zorn, A. M. *et al.* Regulation of Wnt signaling by Sox proteins: XSox17 α/β and XSox3 physically interact with β -catenin. *Mol. Cell* **4**, 487–498 (1999).
24. Shibamoto, S. *et al.* Cytoskeletal reorganization by soluble Wnt-3a protein signalling. *Genes Cells* **3**, 659–670 (1998).
25. Koenig, B. B. *et al.* Characterization and cloning of a receptor for BMP-2 and BMP-4 from NIH 3T3 cells. *Mol. Cell. Biol.* **14**, 5961–5974 (1994).
26. ten Dijke, P. *et al.* Characterization of type I receptors for transforming growth factor- β and activin. *Science* **264**, 101–104 (1994).
27. Schaeren-Wiemers, N. & Gerfin-Moser, A. A single protocol to detect transcripts of various types and expression levels in neural tissue and cultured cells: *in situ* hybridization using digoxigenin-labelled cRNA probes. *Histochemistry* **100**, 431–440 (1993).
28. Hamburger, V. & Hamilton, H. L. A series of normal stages in the development of the chick embryo. *J. Morphol.* **88**, 49–92 (1951).
29. Read, E. M. *et al.* Evidence for non-axial A/P patterning in the nonneural ectoderm of *Xenopus* and zebrafish pregastrula embryos. *Int. J. Dev. Biol.* **42**, 763–774 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank members of the Edlund lab for helpful discussions. We are very grateful to J. Nathans for providing the Wnt8 cell line and mFrz8CRD–IgG plasmid; R. Lovell-Badge and E. Boncinelli for antibodies; C. Healy and M. Kessel for chick *Sox3* and *DLX-5*, respectively; J. C. Hsieh for technical advice; and H. Alstermark for technical assistance. T.E. is supported by the Swedish Medical Research Council and by the Foundation for Strategic Research. T.M.J. is supported by grants from the NIH. T.M.J. and R.N. are Investigators of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to T.E. (e-mail: thomas.edlund@micro.umu.se).

Mae mediates MAP kinase phosphorylation of Ets transcription factors in *Drosophila*

David A. Baker^{*†}, Blandine Mille-Baker^{*}, S. Mark Wainwright[†], David Ish-Horowicz[†] & Nicholas J. Dobb^{*}

^{*} Cell Signalling Unit, Institute of Reproductive and Developmental Biology, Imperial College School of Medicine, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK

[†] Developmental Genetics Laboratory, Imperial Cancer Research Fund, PO Box 123, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

The evolutionarily conserved Ras/mitogen-activated protein kinase (MAPK) cascade is an integral part of the processes of cell division, differentiation, movement and death. Signals received at the cell surface are relayed into the nucleus, where MAPK phosphorylates and thereby modulates the activities of a subset of transcription factors^{1,2}. Here we report the cloning and characterization of a new component of this signal transduction pathway called Mae (for modulator of the activity of Ets). Mae is a signalling intermediate that directly links the MAPK signalling pathway to its downstream transcription factor targets. Phosphorylation by MAPK of the critical serine residue (Ser 127) of the *Drosophila* transcription factor Yan depends on Mae, and is mediated by the binding of Yan to Mae through their Pointed domains. This phosphorylation is both necessary and sufficient to abrogate transcriptional repression by Yan. Mae also regulates the activity of the transcriptional activator Pointed-P2 by a similar mechanism. Mae is essential for the normal development and viability of *Drosophila*, and is required *in vivo* for normal signalling of the epidermal growth factor receptor. Our study indicates that MAPK signalling specificity may depend on proteins that couple specific substrates to the kinase.

Phosphorylation of transcription factors by MAPK is a key link between cell signalling and the control of gene expression³. The Ets family of transcription factors regulates cell growth and differentiation, and the activities of many of its 35 members are modulated through phosphorylation by MAPK⁴. To identify signalling intermediates, we isolated proteins that interact with the *Drosophila* Ets transcription factor Yan. Yan functions downstream of several receptor tyrosine kinase (RTK) pathways, where it regulates the differentiation of many precursor cells^{5–8}. Experiments in tissue-culture cells suggest that Yan functions as a transcriptional repressor whose activity is attenuated rapidly by MAPK phosphorylation⁹.

Using full-length *yan* complementary DNA as a bait, we conducted a yeast two-hybrid screen of an 18-h *Drosophila* embryo library, from which we recovered the full-length cDNA clone shown in Fig. 1a (see Methods). The only region of homology between Yan and the predicted Mae protein is a Pointed (Pnt) domain located at the carboxy terminus (Fig. 1b, c). The Pnt domain (also called the SAM domain) defines a subfamily of Ets proteins, including Ets-1, Ets-2, GABP and TEL from vertebrates, and *Drosophila* Yan and Pnt-P2 (ref. 10). It is also found in Polycomb-group proteins, where it mediates protein–protein interactions, and in MAPK kinase kinases that are components of the MAPK cascade¹¹.

We confirmed the biochemical interaction between Yan and Mae using a fusion of Yan to glutathione S-transferase (GST). Figure 2a shows that Mae binds to Yan (lane 2), and that the Pnt domains of Yan and Mae are essential for this interaction. Thus, mutations that disrupt this domain (Yan Δ (46–107), lane 3; Yan(G84P), lane 4; Mae(G139P), lane 5) abrogate the binding of Mae to Yan. This is the first evidence, to our knowledge, that the Pnt domain of Ets transcription factors mediates a heterotypic protein interaction.

As Yan represses transcription by binding to consensus Ets DNA-binding sites¹², we tested whether Mae interferes with Yan binding to DNA in a gel mobility shift assay. Figure 2c shows that Mae does not bind to the consensus Ets DNA-binding site (lane 2), but it prevents Yan from doing so (compare lanes 3 with 7–9; see Supplementary Information). This inhibition is mediated through the Pnt domain-dependent binding of Mae to Yan (compare lanes 7 and 10). DNA binding by Yan requires a functional Ets DNA-binding domain (Fig. 2c, lanes 3 and 5; see ref. 13), but not intact phosphorylation sites or a functional Pnt domain, because Yan^{ACT} (in which all nine MAPK consensus sites are mutated to alanine⁶) binds the DNA site normally (Fig. 2c, lane 4), as does Yan(G84P) in which an amino acid conserved in all Pnt domains is mutated (Fig. 2c, lane 6).

We next established an assay of transcriptional repression by Yan. Because *Drosophila* S2 cells express *mae* endogenously (data not shown), we expressed the *yan* and *mae* constructs in Cos-7 cells (see Supplementary Information). Yan represses the activity of a TK-luciferase reporter tenfold (Fig. 3a). Repression depends on the Ets DNA consensus site (data not shown), and on the Ets DNA-binding domain in Yan (Fig. 3a). As expected, Mae alone does not affect reporter activity because it lacks an Ets DNA-binding domain (Fig. 3a).

RTK signalling downregulates Yan activity through phosphorylation mediated by Ras and the Erk/Rolled MAPK^{6,9}. Expression of a constitutively active form of Ras (¹²⁵I-Ras) completely abrogates transcriptional repression by Yan, but only in the presence of Mae (Fig. 3b, c). Consequently, Mae is required for Ras to inactivate Yan in Cos-7 cells. An excess of Mae alone can reduce the repressive effect of Yan (~3-fold; Fig. 3b, c), presumably through formation of Mae–Yan heterodimers. However, low quantities of Mae, which alone inhibit the repressive effect of Yan by ~30%, are sufficient to mediate the complete, Ras-dependent inactivation of Yan (Fig. 3c). Furthermore, the non-phosphorylatable Yan^{ACT} mutation is insensitive to Ras/Mae-induced inhibition (Fig. 3b), indicating that Mae inhibits Yan by promoting its phosphorylation by MAPK. Indeed, Yan mutants that cannot bind Mae—Yan(G84P) and Yan Δ (46–107) (see Fig. 3b)—are also resistant to inactivation by Ras/Mae.

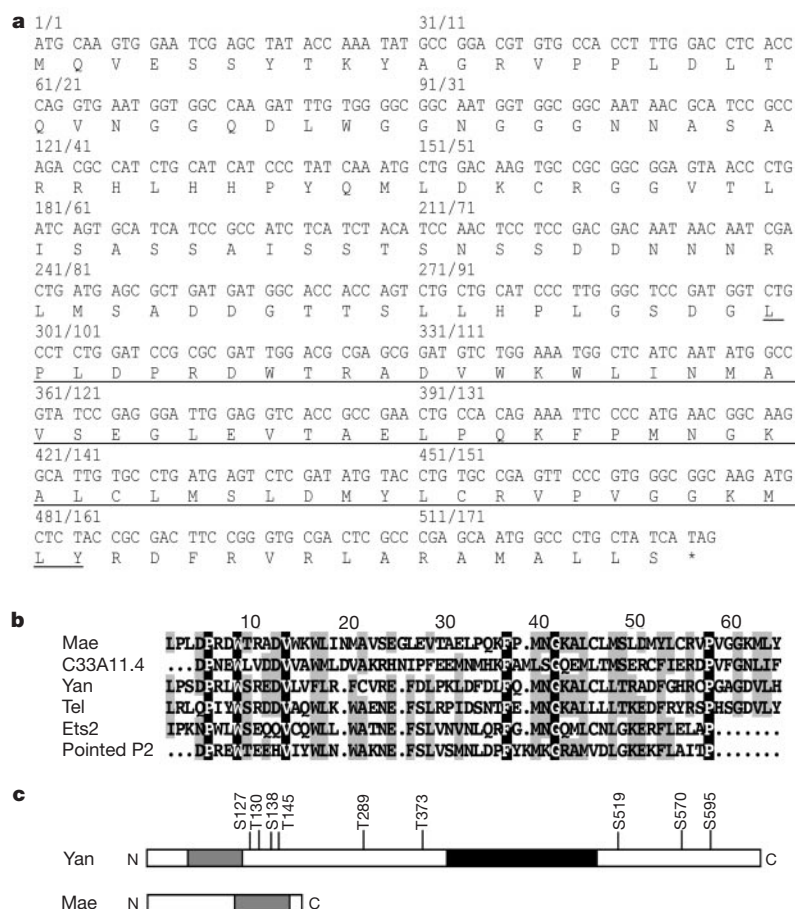


Figure 1 *mae* sequence and gene organization. **a**, The full-length nucleotide sequence and corresponding translation of *mae*. Mae is characterized by the presence of a pointed (Pnt) domain (underlined) and the absence of a DNA-binding Ets domain. **b**, Mae shares significant homology with Pnt domain-containing Ets transcription factors.

c, Representation of Yan and Mae. The Pnt domain is marked as a hatched box at the N terminus of Yan and the C terminus of Mae. The Ets DNA-binding domain of Yan is shown as a black box, and the positions of the nine putative sites of phosphorylation by Erk (PX(S/T/P)) are indicated.

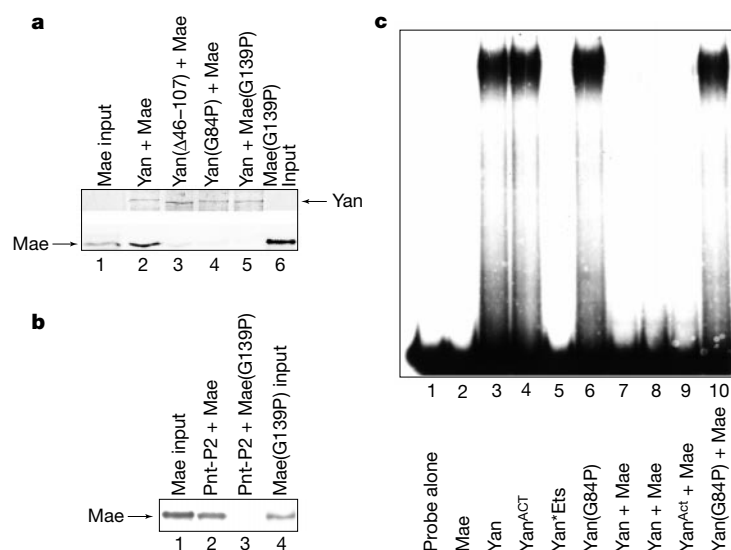


Figure 2 The physical association of Yan and Mae is mediated by their Pnt domains, and this inhibits Yan from binding to DNA. **a**, The indicated GST–Yan fusions were incubated with either ³⁵S-labelled Mae (lanes 2–4) or ³⁵S-labelled Mae(G139P) (lane 5). **b**, Similarly, GST–Pnt-P2 fusions were incubated with either ³⁵S-labelled Mae (lane 2) or Mae(G139P) (lane 5).

c, Gel mobility shift assay. The indicated purified proteins were incubated immediately with the Ets DNA-binding site, except in lane 7 where Yan and Mae were first pre-incubated for 20 min. In lane 8, Yan was first mixed with the Ets DNA-binding site, and then Mae was added for a further 20 min (see Supplementary Information).

Although Yan has nine sites of MAPK phosphorylation (Fig. 1c), Ser 127 is the critical regulatory site of Yan because mutation to alanine creates a constitutive repressor that is refractory to down-regulation by Ras signalling⁶. To test whether Mae mediates phosphorylation of Ser 127 of Yan, we examined *in vitro* phosphorylation of Yan by activated Erk (see Methods). Activated Erk phosphory-

lates wild-type Yan independently of Mae (Fig. 3d, lane 1), and, as expected, is unable to phosphorylate the constitutive repressor Yan^{ACT} (Fig. 3d, lane 2), which has all nine MAPK phosphorylation sites mutated. Notably, Yan(S127), in which all consensus phosphorylation sites except Ser 127 are mutated to alanine, is resistant to phosphorylation by Erk alone (Fig. 3d, lane 4), even with very

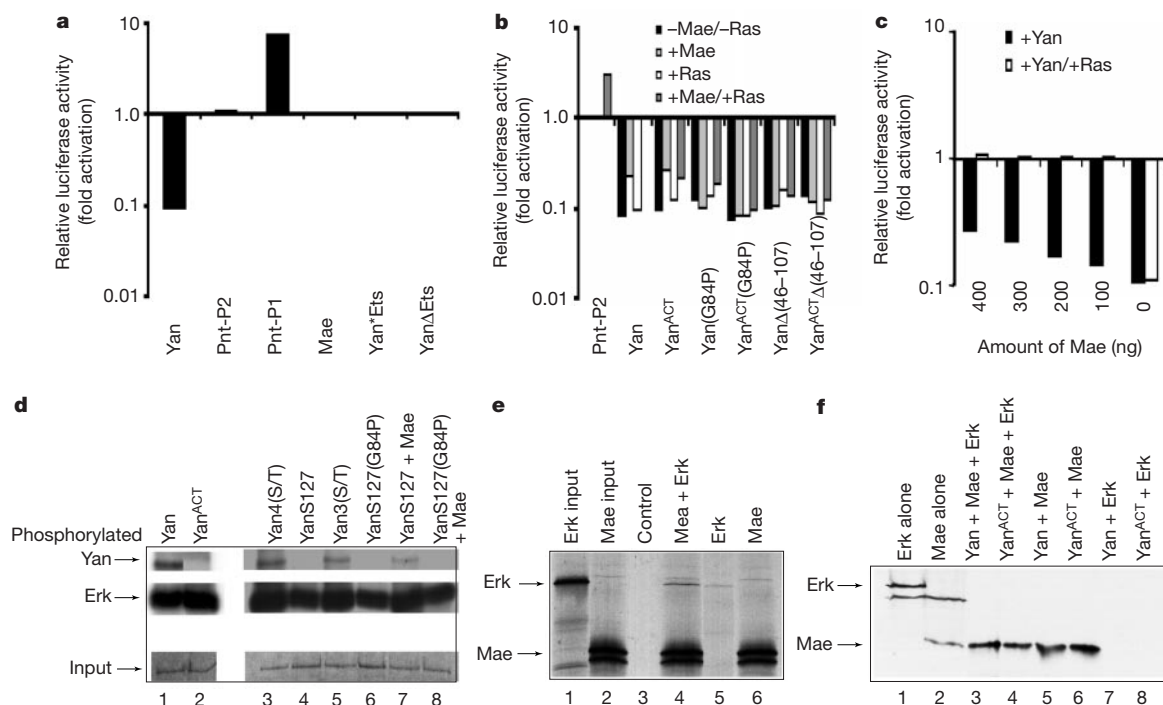


Figure 3 Mae allows Erk kinase to phosphorylate critical serine or threonine residues of Yan and Pnt-P2, and thus regulates Ets transcriptional activity. **a–c**, Luciferase reporter assays of the activity of the Yan and Pnt transcription factors. **a**, Yan is a repressor of transcription, whereas Pnt-P1 is an activator. **b**, **c**, Mae is required for Ras-dependent modulation of the transcriptional activities of Yan and Pnt-P2. **d**, Mae mediates the phosphorylation by Erk of Ser 127 of Yan. The indicated GST-purified proteins were incubated with activated Erk (Calbiochem), with or without Mae. The signal for Yan(S127)

(lane 7) is weaker than that for wild type (lane 1), because it is phosphorylated at a single site. **e**, GST pull-down assay. Erk binding to Yan is weak by comparison with Mae binding to Yan. GST–Yan (lanes 4–6) or GST alone (lane 3) was incubated with the indicated ³⁵S-labelled proteins. **f**, Mae but not Erk co-immunoprecipitates with Yan. Yan-bound Mae and Erk were detected through western blotting with an anti-Myc antibody (lanes 3–8). In lanes 1 and 2, lysates prepared from cells transfected with either *erk* or *mae* were western blotted with an anti-Myc antibody (see Supplementary Information).

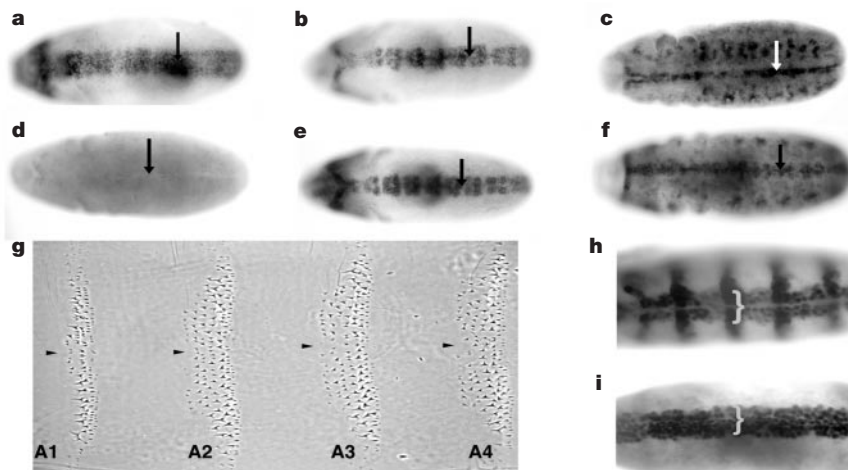


Figure 4 Evidence that *mae* regulates EGFR signalling in the *Drosophila* embryonic neuroectoderm. **a–c**, Ventral views of wild-type embryos at stage 6–7 (**a**), stage 9 (**b**) and stage 11 (**c**), showing *mae* expression along the midline. **d**, Undetectable *mae* expression in *I(2)k06602 / I(2)k06602* embryos (stage 9). **e**, **f**, *I(2)k06602/+* embryos at stage 9 (**e**) and 11 (**f**), showing that the pattern of β -galactosidase expression resembles that of *mae*. The ventral midline is indicated by arrows. The same results were obtained using the

I(2)k12907 mae mutant fly line (data not shown). **g**, Ventral cuticle of *I(2)k12907 / Df(2R)PC4* larvae showing disrupted denticle patterning (arrowheads) around the midline in segments A1–4. **h**, Stage 9–10 *I(2)k12907, mae-lacZ/Cyo, wg-lacZ* embryo stained for β -galactosidase. **i**, Stage 9–10 homozygous *I(2)k12907* embryo (lacking lateral stripes of *wg-lacZ* staining) with broadened *mae* transcription domain. Anterior is to the left. The *mae* transcription domain in **h** and **i** is indicated by a square bracket.

high amounts (up to 1 µg) of the kinase (data not shown). Yan(S127) is efficiently phosphorylated by Erk and Mae together (lane 7), however, indicating that Mae is needed for phosphorylation of this Ser 127 residue.

This phosphorylation depends on the Pnt domain of Yan (Fig. 3d, compare lanes 4 and 7 with lanes 6 and 8), indicating that the ability of Mae to promote Erk-directed phosphorylation of Ser 127 is dependent on binding of Mae to Yan. Yan(S127) represses transcription of the luciferase reporter as efficiently as the wild-type protein, and is also completely susceptible to inactivation by ¹²⁵I-Ras/Mae (see Supplementary Information). Therefore, Mae is required for MAPK to abolish transcriptional repression by Yan, by mediating phosphorylation of the critical regulatory residue Ser 127.

Inhibition of Yan repression by RTK signalling in *Drosophila* is accompanied by activation of a transcription factor encoded by the *pointed* (*pnt*) gene. *pnt* encodes two alternative Ets transcription factors¹⁴: P1, a constitutively active transcriptional activator whose expression is induced by Ras/MAPK signalling; and P2, a transcriptional activator that, like Yan, contains an amino-terminal Pnt domain and whose activity is stimulated through phosphorylation by the Ras/MAPK pathway¹⁵. Activation of Pnt-P2 by Erk also seems to depend on Mae: the two proteins physically associate and this interaction is dependent on the Pnt domain (Fig. 2b, compare lanes 2 and 3). Pnt-P2 is a weak activator in our transcription assay (Fig. 3a) and is not stimulated by Ras alone (Fig. 3b). However, Ras and Mae together augment its transcriptional activity fourfold (Fig. 3b). Thus, Mae also regulates Pnt-P2 activity, presumably by promoting its phosphorylation by MAPK.

These results show that Mae is needed in cultured cells to permit regulation of Ets transcription factor activity by Erk. This requirement was not evident in previous experiments on Yan⁹ as *Drosophila* S2 cells were used in which endogenous *mae* is expressed at high levels (our unpublished data).

To study Mae function *in vivo*, we used *in situ* hybridization to examine its expression in early *Drosophila* embryos. In stage 6–7 embryos, *mae* is expressed in bilaterally symmetric anteroposterior stripes, 3–4-cells wide, that flank the ventral midline (Fig. 4a). These stripes of expression narrow to a width of two cells and one cell adjacent to the midline by stage 9 (Fig. 4b) and stage 11 (Fig. 4c), respectively. *mae* expression is very similar to the expression pattern of epidermal growth factor receptor (EGFR) signalling regulators such as *rhomboid* (*rho*), *vein*, *argos*, *yan* and *pnt*^{16–20}, and it marks the ventral neurectodermal zone that is patterned by EGFR signalling. *mae* is subsequently expressed in tracheal pits, in ventral denticle domains, and in areas such as the optic lobe and the medial domains of the brain (data not shown), all of which are sites of EGFR activity²¹.

To investigate further the *in vivo* role of *mae*, we analysed the enhancer-trap lines *l(2)k06602* (ref. 22) and *l(2)k12907* (ref. 23), which contain single P-element insertions that are responsible for the lethal phenotype²³. We found that these P-elements insert into the 5' untranslated region (UTR) of the *mae* gene (0.8-kilobases (kb) and 50-base-pairs (bp) upstream of the *mae* initiation codons in *l(2)k06602* and *l(2)k12907*, respectively; see Supplementary Information). β-Galactosidase in early *l(2)k06602* and *l(2)k12907* embryos is, like *mae*, expressed in the ventral neurectoderm, tracheal pits and ventral denticle belts (see Fig. 4e, f). Furthermore, homozygous *mae* mutant embryos (~25% of embryos derived from either heterozygous *l(2)k06602* or *l(2)k12907* parents) lack detectable *mae* expression (see Fig. 4d). These results show that the P-element insertions affect development because they disrupt *mae* expression.

A role for Mae in EGFR signalling is further supported by certain features of the *mae* mutant phenotype. Embryonic patterning is affected in *mae* mutant embryos, especially towards the midline of the anterior abdominal segments. Ventral denticle belts are

thinner, and rows of denticles, particularly the first, are missing or misorientated (Fig. 4g). *mae*^{k12907} and *mae*^{k06602} embryos each show a similar phenotype when hemizygous for *Df(2R)PC4*, a deficiency for the region, indicating that the patterning defects result from lack of *mae* expression. Preferential disruption of midline patterning in anterior abdominal segments is reminiscent of the phenotype of *rho* and *pnt* mutant larvae^{24,25}.

mae seems to be a downstream component of the EGFR signalling pathway. Its pattern of expression depends on Rho, the spatially restricted limiting component of EGFR signalling²⁶, but is not required for *rho* expression (D.A.B., unpublished data). Furthermore, in the *l(2)k12907* enhancer-trap line in which *mae* expression is abolished, *lacZ* marks the *mae* transcriptional domains (and thereby domains of EGFR signalling), and this domain is broadened along the ventral midline (Fig. 4h, i). Similarly, the expression domain of *argos* is broadened in early stage *mae* mutant embryos, although levels of expression, notably in later stage embryos, seem to be reduced (data not shown). *argos* is a downstream target of EGFR signalling whose expression is also disrupted in *Drosophila* embryos that are mutant for regulators of EGFR signalling such as Ras, Yan, Pnt^{18,27,28} and Rho (D.A.B., unpublished data).

The above experiments show that Mae mediates inactivation of Yan by MAPK phosphorylation of the critical regulatory Ser 127 residue. Presumably, binding of Mae causes a local change in the conformation of Yan, which exposes Ser 127 to MAPK. Further evidence for this idea is that Mae binding to the Pnt domain in Yan also affects the ability of Yan to bind DNA through its Ets domain (Fig. 2b). Erk can associate with Mae and Yan in GST pull-down assays but, unlike Mae binding to Yan, this interaction is weak and is not evident in co-precipitation assays (Fig. 3e, f; and Supplementary Information). We propose that Mae binding to Yan or Pnt-P2 allows MAPK to phosphorylate their critical residues in a transitory ternary enzyme–substrate complex.

To our knowledge, this is the first report of a requirement for an intermediary protein linking MAPK to its substrate, and it suggests a mechanism for achieving tissue-specific responses to the generic RTK/Ras/MAPK signalling pathways; that is, determination of the MAPK substrate through tissue-specific adapter/coupling proteins. Cells will respond differently to MAPK according to the adapter/coupling proteins that they express. Future work will help test whether phosphorylation of other MAPK substrates also depends on adapter/coupling proteins. □

Methods

Yeast two-hybrid and cDNA cloning

Full-length *yan* cDNA was cloned into pAS2-1 (Clontech) in-frame with the GAL4 DNA-binding domain. Independent transformants (5×10^5) from a GAL4 activation domain *Drosophila* embryo library were screened in Hf7c yeast cells according to the manufacturer's protocol (Clontech). We confirmed the specificity of interaction by transforming yeast with the cDNA of positive clones, in either the presence or the absence of the *yan* bait or an unrelated bait. A partial *mae* cDNA was isolated and used to screen a λGT11 1–18 h *Drosophila* embryo cDNA library for a complete *mae* cDNA.

Molecular biology and biochemistry

We performed mutagenesis using the Transformer site-directed mutagenesis kit (Clontech). For pull-down assays, Mae proteins were labelled with [³⁵S]methionine using the TNT-coupled reticulocyte *in vitro* translation system (Promega), and incubated with GST–Yan fusions that were immobilized on glutathione–sepharose beads. For gel mobility shift assays, proteins were incubated with 0.1 ng of ³²P-labelled probe for 20 min at 4 °C in the presence of 2 µg of poly(dI/dC), 110 mM KCl, 10 mM HEPES, 5.5 mM MgCl₂, 5 mM sodium β-glycerophosphate, 0.05 mM EDTA, 0.05 mM spermine and 17.5% glycerol pH 7.6.

In vitro kinase assays were performed in 'kinase' buffer (10 µCi [³²P]ATP, 25 mM sodium β-glycerophosphate, 20 mM MOPS pH 7.2, 10 mM MgCl₂, 10 mM MnCl₂, 2 mM NaF, 1 mM dithiothreitol and 1 mM NaVO₄) for 30 min at 30 °C, and were resolved by SDS–PAGE. For co-immunoprecipitation, Yan was fused in-frame with the Flag epitope, and Mae and Erk were each fused in-frame with the Myc epitope. We immunoprecipitated Yan with anti-Flag antibody (Sigma) from lysates prepared from Cos-7 cells (1% Nonidet P40, 10% glycerol, 75 mM NaCl, 20 mM Tris pH 7.5) that had been transfected with the appropriate constructs. Associated proteins were eluted and resolved using SDS–PAGE. Western blotting was carried out as described²⁹.

Luciferase reporter assays

Constructs were transfected using the calcium-phosphate method into 1×10^5 Cos-7 cells growing in 2.4-cm wells in DMEM without serum. After a single change of medium 7 h after transfection, cells were lysed for analysis of luciferase activity after 16 h of culture. We used 1 µg of reporter plasmid and 400 ng of all other constructs. The total amount of DNA to be transfected was adjusted using control vector. Equal transfection efficiencies were determined by co-transfection of the control pRL-TK vector and assay with Renilla luciferase (Promega). In every experiment, each condition was examined in triplicate.

Drosophila genetics and immunohistochemistry

Stocks were obtained from the Bloomington Stock Centre, G. Rubin and M. Freeman. *mae* stocks were maintained using balancer chromosomes expressing green fluorescent protein or *wg-lacZ* to allow selection of homozygous mutants. Cuticle preparations of mutant larvae were done as described³⁰. Whole-mount *in situ* hybridization was done as described³¹ using digoxigenin-labelled (Roche) anti-sense RNA probes. Probes were derived from the following cDNAs: the 3'-most 800 bp of *mae* including 250 bp of the 3' UTR; a 1.1-kb 5' *Bam*H1 fragment of *yan* that excluded the *ets* DNA-binding domain; the coding region of *argos*; a 2.5-kb fragment of *rho* encompassing the 1-kb coding region. *lacZ* expression was detected using rabbit anti-β-galactosidase antibodies (Kappel) and alkaline-phosphatase coupled anti-rabbit antibodies (Jackson Labs).

Received 7 February; accepted 16 March 2001.

- Hill, C. S. & Treisman, R. Transcriptional regulation by extracellular signals: mechanisms and specificity. *Cell* **80**, 199–211 (1995).
- Marshall, C. J. Specificity of receptor tyrosine kinase signalling: transient versus sustained extracellular signal-regulated kinase activation. *Cell* **80**, 179–185 (1995).
- Treisman, R. Regulation of transcription by MAP kinase cascades. *Curr. Opin. Cell Biol.* **8**, 205–215 (1996).
- Wasylyk, B., Hagman, J. & Gutierrez-Hartmann, A. Ets transcription factors: nuclear effectors of the Ras-MAP-kinase signaling pathway. *Trends Biochem. Sci.* **23**, 213–216 (1998).
- Lai, Z. C. & Rubin, G. M. Negative control of photoreceptor development in *Drosophila* by the product of the *yan* gene, an ETS domain protein. *Cell* **70**, 609–620 (1992).
- Rebay, I. & Rubin, G. M. Yan functions as a general inhibitor of differentiation and is negatively regulated by activation of the Ras/MAPK pathway. *Cell* **80**, 857–866 (1995).
- Rogge, R. *et al.* The role of yan in mediating the choice between cell division and differentiation. *Development* **121**, 3947–3958 (1995).
- Rebay, I. *et al.* A genetic screen for novel components of the Ras/mitogen-activated protein kinase signalling pathway that interact with the *yan* gene of *Drosophila* identifies split ends, a new RNA recognition motif-containing protein. *Genetics* **154**, 695–712 (2000).
- O'Neill, E. M., Rebay, I., Tjian, R. & Rubin, G. M. The activities of two ETS-related transcription factors required for *Drosophila* eye development are modulated by the Ras/MAPK pathway. *Cell* **78**, 137–147 (1994).
- Jousset, C. *et al.* A domain of TEL conserved in a subset of ETS proteins defines a specific oligomerization interface essential to the mitogenic properties of the TEL-PDGFRβ oncoprotein. *EMBO J.* **16**, 69–82 (1997).
- Ponting, C. P. SAM: a novel motif in yeast sterile and *Drosophila* polyhomeotic proteins. *Protein Sci.* **4**, 1928–1930 (1995).
- Nye, J. A. Interaction of murine ETS-1 with GGA-binding sites establishes the ETS domain as a new DNA-binding motif. *Genes Dev.* **6**, 975–990 (1992).
- Soudant, S. *et al.* A residue in the ETS domain mutated in the v-ets oncogene is essential for the DNA-binding and transactivating properties of the ETS-1 and ETS-2 proteins. *Nucleic Acids Res.* **22**, 3871–3879 (1994).
- Klamt, C. The *Drosophila* gene *pointed* encodes two ETS-like proteins which are involved in the development of midline glial cells. *Development* **117**, 163–176 (1993).

- Brunner, D. *et al.* The ETS domain protein Pointed-P2 is a target of MAP kinase in the sevenless signal transduction pathway. *Nature* **370**, 386–389 (1994).
- Golembo, M., Raz, E. & Shilo, B.-Z. The *Drosophila* embryonic midline is the site of Spitz processing, and induces activation of the EGF receptor in the ventral ectoderm. *Development* **122**, 3363–3370 (1996).
- Yagi, Y., Suzuki, T. & Hayashi, S. Interaction between *Drosophila* EGF receptor and *vnd* determines three dorsoventral domains of the neuroectoderm. *Development* **125**, 3625–3633 (1998).
- Gabay, L. *et al.* EGF receptor signalling induces *pointed P1* transcription and inactivates Yan protein in the *Drosophila* embryonic ventral ectoderm. *Development* **122**, 3355–3362 (1996).
- Golembo, M., Yarnitzky, T., Volk, T. & Shilo, B.-Z. Vein expression is induced by the EGF receptor pathway to provide a positive feedback loop in patterning the *Drosophila* embryonic ventral ectoderm. *Genes Dev.* **13**, 158–162 (1999).
- Gabay, L., Seger, R. & Shilo, B.-Z. *In situ* activation pattern of *Drosophila* EGF receptor pathway during development. *Science* **277**, 1103–1106 (1997).
- Dumstrei, K. *et al.* EGFR signalling is required for the differentiation and maintenance of neural progenitors along the dorsal midline of the *Drosophila* embryonic head. *Development* **125**, 3417–3426 (1998).
- Roch, F. *et al.* Screening of larval/pupal P-element induced lethals on the second chromosome in *Drosophila melanogaster*: clonal analysis and morphology of imaginal discs. *Mol. Gen. Genet.* **257**, 103–112 (1998).
- Spradling, A. C. *et al.* The Berkeley *Drosophila* genome gene disruption project. Single P-element insertions mutagenizing 25% of vital *Drosophila* genes. *Genetics* **153**, 135–177 (1999).
- Jurgens, G., Wieschaus, E., Nusslein-Volhard, C. & Kluding, H. Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster* II. Zygotic loci on the third chromosome. *Roux's Arch. Dev. Biol.* **193**, 283–295 (1984).
- Mayer, U. & Nusslein-Volhard, C. A group of genes required for pattern formation in the ventral ectoderm of the *Drosophila* embryo. *Genes Dev.* **2**, 1496–1503 (1988).
- Wasserman, J. D., Urban, S. & Freeman, M. A family of rhomboid-like genes: *Drosophila rhomboid-1* and *roughoid/rhomboid-3* cooperate to activate EGF receptor signalling. *Genes Dev.* **14**, 1651–1663 (2000).
- Golembo, M., Schweitzer, R., Freeman, M. & Shilo, B.-Z. *Argos* transcription is induced by the *Drosophila* EGF receptor pathway to form an inhibitory feedback loop. *Development* **122**, 223–230 (1996).
- Scholz, H., Sadlowski, E., Klaes, A. & Klamt, C. Control of midline glia development in the embryonic *Drosophila* CNS. *Mech. Dev.* **64**, 137–151 (1997).
- Towbin, H., Staehelin, T. & Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl Acad. Sci. USA* **76**, 4350–4354 (1979).
- Wieschaus, E. & Nusslein-Volhard, C. in *Drosophila, a Practical Approach* (ed. Roberts, D. B.) 199–226 (IRL, Oxford, 1986).
- Tautz, D. & Pfeifle, C. A non-radioactive *in situ* hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene *hunchback*. *Chromosoma* **98**, 81–85 (1989).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

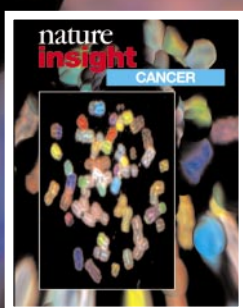
Acknowledgements

We thank S. Bullock, R. Winston, H. McNeill, P. Verrijzer, C. Hill, A. Maata, P. Mason, J. White, J. Brossens, M. Freeman, G. Rubin, T. Laverty and P. Soccorso for advice, discussions and gifts of materials. This work was supported by the Imperial Cancer Research Fund and BBSRC.

Correspondence and requests for materials should be addressed to D.A.B. (e-mail: d.baker@ic.ac.uk). The sequence has been deposited in GenBank under accession number AF358670.

nature insight

Cancer



Cover illustration

Frequent occurrence of chromosomal aberrations visualized by spectral karyotyping in human lymphocytes irradiated with 4 Gy of ionizing radiation. (Image courtesy of H. B. Beverloo and J. J. Boei, MGC, Rotterdam/Leiden.)

Cancer is an umbrella term covering a plethora of conditions characterized by unscheduled and uncontrolled cellular proliferation. As the average age in many countries steadily rises, so do cancer-related deaths, so that cancer will be one of the most common causes of death in the 21st century. Almost any mammalian organ and cell type can succumb to oncogenic transformation, giving rise to a bewildering array of clinical outcomes.

The causes of cancer are many and varied, and include genetic predisposition, environmental influences, infectious agents and ageing. These transform normal cells into cancerous ones by derailing a wide spectrum of regulatory and downstream effector pathways. It is just this complexity that has hampered the development of effective and specific cancer therapies.

Any attempt to provide a comprehensive overview of cancer-related knowledge would be futile — there are around 1.3 million cancer-related *Medline* entries. We have therefore focused on topics undergoing particularly rapid progress, and aimed to provide a balanced picture of the diverse disciplines associated with cancer research. The articles represent particular highlights selected by the editors and authors. Exclusion of important science does not constitute a value judgement.

Ponder leads this collection on page 336 with an article on cancer genetics, highlighting the importance of cell-external and epigenetic events on cancer development. The main hallmark of cancer is deregulated cellular proliferation, and Evan and Vousden provide an overview of cell-cycle control and apoptosis on page 342. Two papers survey signal transduction from the cell surface to targets mediating the proliferative response. On page 349, Beachy and Taipale concentrate on the Hedgehog and Wnt signalling pathways, while Blume-Jensen and Hunter's review centres on deregulation of tyrosine kinase signalling on page 355. Cancer incidence would be very much higher without efficient damage-repair pathways, and Hoeijmakers reviews the main mechanisms of genomic maintenance on page 366. Liotta and Kohn go outside the cell on page 375 and describe how the tumour microenvironment affects cancer progression. On page 380, Rosenberg discusses tumour immunology and the promising potential of immunotherapy, while Appelbaum approaches immunotherapy from the angle of bone-marrow transplantation on page 385. Finally, Peto provides striking insights into cancer epidemiology on page 390.

A further port of call for those interested in cancer-related research is *Nature Reviews Cancer* (prelaunch website now live at www.nature.com/reviews/cancer). We are pleased to acknowledge the financial support of AstraZeneca in producing this Insight. As always, *Nature* carries the sole responsibility for all editorial content and peer-review.

We hope that you will share our excitement on reading these articles, which epitomize how diverse and dynamic cancer research is at the start of the 21st century.

Bernd Pulverer Senior Editor

Lesley Anson, Chris Surridge Insight Programme Editors
Liz Allen Publisher

Cancer genetics

Bruce A. J. Ponder

CRC Department of Oncology, University of Cambridge, Hutchison/MRC Research Centre, Hills Road, Cambridge CB2 2XZ, UK
(bajp@mole.bio.cam.ac.uk)

Cancer genetics has for many years focused on mutational events that have their primary effect within the cancer cell. Recently that focus has widened, with evidence of the importance of epigenetic events and of cellular interactions in cancer development. The role of common genetic variation in determining the range of individual susceptibility within the population is increasingly recognized, and will be addressed using information from the Human Genome Project. These new research directions will highlight determinants of cancer that lie outside the cancer cell, suggest new targets for intervention, and inform the design of strategies for prevention in groups at increased risk.

With few exceptions, cancers are derived from single somatic cells and their progeny. The cells in the emerging neoplastic clone accumulate within them a series of genetic or epigenetic changes that lead to changes in gene activity, and so to altered phenotypes which are subject to selection¹. Ultimately, a cell population evolves that can disregard the normal controls of proliferation and territory and become a cancer. Hanahan and Weinberg² identify six 'hallmark features' of the cancer cell phenotype: disregard of signals to stop proliferating and of signals to differentiate; capacity for sustained proliferation; evasion of apoptosis; invasion; and angiogenesis.

Several factors can influence the evolution of a cancer. They are summarized in Fig. 1. The bold horizontal arrows represent the pathway of successive genetic or epigenetic events through which the cell acquires the cancer phenotype. Mostly these are somatic events, but in many of the inherited cancer syndromes, discussed below, one of the events is inherited. The alternative pathways to the right signify that overtly similar cancers may contain different combinations of genetic events, which may confer different properties. (This is the basis of 'molecular profiling' of tumours to predict clinical behaviour³.)

Influences on the pathway are represented as vertical arrows. One set of influences affects the probability that a pathway event will occur. Within the cancer cell these include acquired or inherited defects in DNA repair or in cell-cycle checkpoints (see articles in this issue by Evan and Vousden, pages 342–348, and Hoeijmakers, pages 366–374), and, possibly, defects in the regulation of epigenetic events⁴. The production and destruction of endogenous mutagens such as free radicals will also affect the probability of mutational events, and may be modified by genetic variation. External influences include environmental exposures, for example diet or cigarette smoke, the response to which again may be modified by genetic variation in metabolic systems acting inside or outside the cell (refs 5, 6, and see article in this issue by Peto, pages 390–395).

Other factors influence the outcome of pathway events once they have occurred. Within the cell, these might be any type of variation that modifies the effect of the pathway event on the cellular phenotype, or the response of the altered cells to signals from outside. Outside the cell, possible influences include paracrine interactions with neighbouring cells⁷ and systemic effects such as the effectiveness of cellular defence mechanisms against the developing cancer, or levels of

circulating hormones or growth factors^{8,9}. Normal genetic variation in these factors is likely to be the source of much of the low-level predisposition to cancer, and of the genetic modifier effects seen in human and experimental tumours^{10,11}.

Before focusing on the factors that influence carcinogenesis, we should first consider the historical development of ideas surrounding events on the main pathway of cancer development.

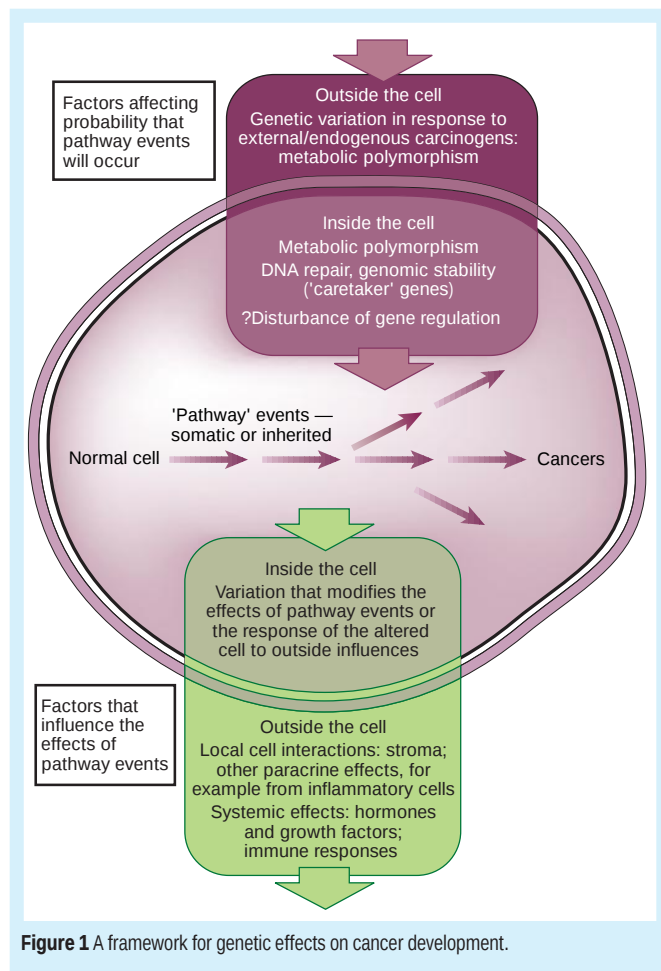
Events on the cancer pathway

The idea that tumours arise from somatic genetic change originated in the early 1900s. It was not until the necessary technologies became available in the early 1970s that tumour formation was related to the action of specific genes. The concepts that developed were of course shaped by the assays on which they were based. The idea of gain-of-function genetic alterations came from experiments that involved gene transfer into recipient cells; these cells could then be assayed for 'transformation' — an approximation to a cancer phenotype. The idea of loss-of-function genetic change came from two different directions: from epidemiology and the study of inherited predisposition¹², and from cell hybridization experiments in which malignancy was found to be recessive to the normal phenotype¹³. This history is relevant because, even today, our partial knowledge of the development of cancer is necessarily constrained by the assays we have available.

Gain-of-function genetic events

The key concept in relation to gain-of-function events is the 'oncogene' (for review, see ref. 14). By the late 1960s, it had been shown that cells in culture could be transformed by several DNA viruses and retroviruses, and subsequently that a single gene from these viruses (the first example was *src*, from Rous sarcoma virus) could carry out this transformation. Genes related in sequence to those in the transforming retroviruses were found in the DNA of normal cells; these genes had functions in the control of normal cell growth or differentiation, but their inappropriate activation by a variety of mechanisms could lead to cancer. The normal cellular genes were termed 'proto-oncogenes'; their activated counterparts were 'oncogenes'¹⁵.

In the late 1970s, fragmented DNA from human cancer cells was transferred into cultured non-neoplastic cells (mouse NIH-3T3 fibroblasts were used) by transfection. The aim was to see if transformation would result and, if so, to recover the active DNA sequences. The first transforming gene to be recovered from human cancer cells by this technique turned out to



be a mutant form of *Ha-ras*, a proto-oncogene already familiar from retroviral studies¹⁶. Similar experiments have since identified many more transforming oncogenes¹⁴, although these probably reflect only a subset of all gain-of-function genetic changes in cancer cells. For example, not all cells are good recipients for transfection, and the predominant use of rodent fibroblasts and of assays for 'transformation', rather than other aspects of the cellular phenotype that might be relevant, may have restricted the range of genes that could be found.

A further line of evidence for the role of activation of specific genes in cancer came from better techniques of chromosome analysis, starting with chromosome banding in the 1970s. In some tumours there were chromosomal translocations with consistent breakpoints, and some of these breakpoints proved to be in, or near to, already described proto-oncogenes — for example, *c-myc* in Burkitt's lymphoma and *c-abl* in chronic myelogenous leukaemia¹⁷. In others, there were consistent regions of chromosomal amplification¹⁸. The inference, borne out by experiment, was that these specific chromosomal events could result in increased expression or activity of the related genes. Many more examples have been found¹⁹, predominantly in haematological cancers and sarcomas where chromosomal identification is technically straightforward. A current question is whether the recently introduced techniques of chromosome analysis by molecular hybridization^{20,21} will reveal similar mechanisms among the more complex chromosomal changes in epithelial malignancies, or whether perhaps epithelial cancers have different genetic mechanisms of development²².

Loss-of-function genetic events

The impression given by the gene-transfer studies is of a single-step, gain-of-function mechanism for carcinogenesis. But this is a bias imposed by the methods used. The first evidence for loss-of-function

genetic changes came from studies of children's cancers, in particular retinoblastoma. Like many cancers, retinoblastoma occurs in an inherited and a sporadic form. Knudson¹² described the distribution of age at diagnosis in inherited and sporadic cases. In inherited cases the distribution was consistent with a requirement for one further event for tumour formation. This event occurred with constant probability over time. In sporadic cases, the age distribution was more complex, and consistent with a need for two events. The inference was that in either case, two rate-limiting events were needed to form the tumour, and that in inherited cases one of these was already present in the germline. Comings²³ suggested that the two events might affect the two alleles of a single gene, implying that their effects would be recessive at the cellular level. Subsequently, in some inherited cases, a germline deletion was found on chromosome 13, implying that loss of a gene in that region might be the first event. This led to biochemical and molecular studies which showed that tumour development did indeed require loss of both copies of that region of chromosome 13 (ref. 24); using the chromosomal deletions as signposts, the *Rb* gene was ultimately cloned and found to be mutated in both copies in the tumours. *Rb* is thus the prototype of the class of 'tumour-suppressor genes'²⁵ where, in distinction to oncogenes, loss of function is required for tumorigenesis.

Linkage and positional cloning in inherited cancer syndromes has identified many more tumour-suppressor genes (for review, see ref. 26). Loss-of-function mutations are much more common than gain-of-function mutations in inherited predisposition, presumably because the loss of function is masked by the remaining normal allele during development (except in the recessive DNA-repair deficiencies), whereas a gain-of-function cancer-promoting mutation might well be lethal. In most inherited cancers, the germline loss-of-function allele represents one step on the pathway shown in Fig. 1, and in most cases, as in retinoblastoma, the same genes are involved by somatic mutation in non-hereditary forms of the same cancer.

If the definition of a 'tumour-suppressor gene' were only that loss of function should contribute to cancer, then a list of potential genes could include not only genes such as *Rb*, but also a wider variety of genes acting at different points in Fig. 1. One might, for example, include genes that determine skin pigmentation as suppressors, on the grounds that fair-skinned individuals have a higher risk of skin cancer. Used as broadly as this, the term is perhaps of little help. Haber and Harlow²⁷ suggested a tighter definition which required the unequivocal demonstration of inactivating mutations of the gene. This had a practical rather than conceptual purpose — to lay down some unambiguous criteria by which the validity of the numerous candidates proposed as new suppressor genes could be judged. But four years later, we might be concerned that the requirement for mutation excludes genes where the predominant mechanism of loss of function is epigenetic⁴. If the term 'suppressor' is restricted to genes whose action lies within the cancer cell, two categories may usefully be distinguished. The first contains genes like *Rb* whose loss of function (by whatever mechanism) is rate limiting for cancer development and which lie on the direct pathway shown in Fig. 1 — the 'classical' tumour suppressors, termed 'gatekeepers' by Kinzler and Vogelstein²⁸. Cancer predisposition due to these genes is tissue specific, although the mechanism of the specificity is generally unclear. The second group contains genes whose loss of function accelerates the acquisition of pathway events, but whose loss is not essential, and whose action lies outside the pathway itself. These are genes involved in DNA repair and genome integrity, which have been termed 'caretakers'²⁸. (For details of DNA-repair genes, see review in this issue by Hoeijmakers, pages 366–374).

Somatic loss of a suppressor gene allele often involves a loss of chromosomal material, ranging in extent from a sub-band to the whole chromosome. Such events are conveniently assayed by 'loss of heterozygosity' (LOH), which is a comparison of polymorphic loci in DNA from blood and tumour in the same individual, and the finding of contiguous regions of tumour DNA where one allele is absent.

These regions might be expected to contain suppressor genes. LOH analysis has identified large numbers of regions of chromosomal loss in many of the common cancers²⁹, but the number of suppressor genes that have been identified convincingly, by the criterion of somatic mutation in the remaining allele, is small. There are several possible explanations: most LOH are noise; they reflect haploinsufficiency³⁰; or perhaps the mutational criterion for identifying a suppressor gene is too stringent. In particular, there is growing evidence that epigenetic silencing rather than mutation is a common mechanism for loss of suppressor gene function.

Epigenetic pathway events

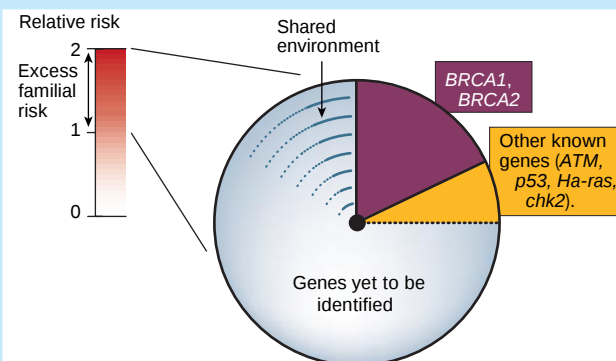
Epigenetic regulation of gene expression by methylation is an important mechanism of the determination of cell fate in embryogenesis. Disturbance of epigenetic mechanisms in the special case of genomic imprinting are responsible, for example, for loss of imprinting (LOI) and hence overexpression of the gene encoding insulin-like growth factor (IGF)-2 in the pathogenesis of Wilms tumour in Beckwith–Weidemann syndrome³¹, and in some epithelial cancers, including colonic cancer³². It has been shown that methylation of regions rich in cytosine–guanine doublets ('CpG islands') in the promoter region in somatic cells is a common mechanism of epigenetic silencing of one or sometimes both alleles of tumour-suppressor genes such as *VHL*, *mlh1*, *p16* (*CDKN4/p16^{INK4A}*) and possibly *BRCA1* (ref. 4). It is not clear whether the epigenetic silencing of particular genes in cancer occurs through a stochastic process followed by selection, or whether certain promoters are predisposed (and if so, what might be the mechanisms involved) (reviewed in ref. 4). It is also unclear what determines whether a particular gene will lose function by an epigenetic or a mutational mechanism. Loss of function of the cyclin-dependent kinase inhibitor *p16* may occur through deletion, point mutation or promoter hypermethylation, but the frequency of each mechanism differs between tumour types⁴. Within the same tumour type, the mechanism may differ in different contexts. Thus, germline mutation of the *MLH1* gene is frequent in familial colon cancers with the microsatellite instability phenotype; but in sporadic cancers with this phenotype, promoter hypermethylation and loss of expression of *MLH1* (and, interestingly, LOI of the IGF-2 gene) is more common³³.

Although promoter hypermethylation has clearly been implicated in silencing of suppressor genes, there are other mechanisms by which changes in methylation might contribute to tumorigenesis. Examples come from the experimental manipulation of the activity of the maintenance DNA methylase *Dnmt1* in mice. Thus, there is a reported increase in somatic mutation in mice heterozygous for loss of function of *Dnmt1*³⁴, and for widespread changes in gene expression in *Dnmt1*^{-/-} mouse embryo fibroblasts rescued from apoptosis by inactivation of *p53* (ref. 35). The reduced incidence of intestinal adenomas in *Min* mice heterozygous for a *Dnmt1*-null allele³⁶ (which seems counter to the increase in somatic mutation reported above) indicates that changes in genomic methylation may modify the phenotypic expression of a strong predisposing gene. The mechanisms of these effects, and their relevance to human cancer, require further investigation.

Epigenetic mechanisms can lead to a progressive, although patchy, silencing of some genes with age³⁷. It is interesting to speculate to what extent our tissues may be a progressive mosaic either of gene silencing (or in the case of the IGF-2 gene, for example, of loss of imprinting), and what factors might influence this process³⁸. The progressive silencing with age of the expression of β -galactosidase reporter genes in transgenic mice is well known. This is a highly artificial experimental situation which may have no relevance at all to endogenous genes in human tissues. Nevertheless, it is intriguing that histochemical staining of tissue sections showed the β -gal expression often to be strongly mosaic in intensity; the size of the positive patches diminished with age, and both the grain of the mosaic and its rate of disappearance differed on different genetic

Box 1

What genes might account for familial breast cancer?



Averaged across all ages, the risk of breast cancer to the sister, mother or daughter of a case is increased about twofold, as illustrated in the figure above. This excess familial risk provides an upper estimate (assuming all the risk is genetic) of the genetic effect that must be explained.

Modelling of genes that might be involved

The relationship between the familial relative risk (FRR) and the frequency and strength of predisposition of any predisposing allele is given by $FRR = [1 + p(1-p)(RR-1)^2] / [1 + p(RR-1)]^2$ where p is the allele frequency, and RR is the cancer risk in a carrier versus a non-carrier of the allele.

Assuming for purposes of illustration that the predisposing alleles are dominant, the table below shows some worked examples of the types of genetic effect that might explain the remaining familial clustering of breast cancer once *BRCA1* and *BRCA2* are accounted for. (Note that the real situation is quite unknown.)

	RR	Frequency of predisposing alleles in population	Contribution to excess familial risk	Number of such genes needed to account for all the observed familial risk*
'BRCA3'-like	10	0.002	0.16	4–5
	1.5	0.01	0.0025	240–320
Common, low-penetrance genes	1.5	0.1	0.020	30–40
	1.5	0.3	0.045	13–18

*Lower figure assumes that genes combine multiplicatively, upper figure assumes that genes combine additively.

backgrounds³⁹. Crosses between the relevant strains mapped a controlling locus to mouse chromosome 4 (ref. 40). Inheritance of methylation patterns in human DNA has also been described⁴¹. It is possible that susceptibility to cancer may be influenced by inherited variation in genes that regulate epigenetic silencing.

Patterns of pathway events

It has been estimated that between four and seven rate-limiting genetic events are needed for the development of the common epithelial cancers⁴². Because, presumably, the constraints to be overcome vary in importance between tissues, and can be evaded in different ways (for example, a signalling pathway may be disrupted at different points; see articles in this issue by Blume-Jensen and Hunter, pages 355–365, and Taipale and Beachy, pages 349–354), it is not surprising that the precise pattern of genetic alterations differs between cancers of different types, and of the same type^{3,43}. But the patterns are not random. Specific associations of events are seen within individual tumours, and these presumably reflect the evolution of the tumours along particular pathways, as suggested in Fig. 1. Such patterns might potentially be important in several practical ways. They are the basis for the current optimism that 'molecular

Table 1 Inherited predisposition to cancer

	Contribution to overall cancer incidence	Clinical features	Frequency of predisposing alleles	Effect on individual risk
Inherited cancer syndromes	1–2% at most	Rare/unusual cancers or combinations of cancers. Sometimes with associated developmental defects or non-neoplastic phenotype. Mendelian dominant inheritance	Rare (~1:1,000 or less)	Strong: lifetime risks of cancer up to 50–80%
Familial cancers	?Up to 10% depending on definition	Families with several cases of common cancers that fall into a recognized pattern of cancer types (for example, breast and ovary; colon+endometrium+urinary). Spectrum from families with multiple cases at young age (strongest evidence of predisposition) to two or three cases at older ages: many of the latter will be due to chance or to combinations of weaker genes. Generally show pattern consistent with dominant inheritance.	Uncommon to common	Moderate to weak
Predisposition without evident family clustering	No precise figure possible. Distribution of risk within population may result in substantial fraction of cancer incidence within predisposed minority	Single cases of cancer at any site, some with one or two affected relatives. The distribution of these cases in the population is probably determined by the combined effects of multiple genetic and non-genetic risk factors.	Multiple common alleles	Weak

profiling' of tumours by genomic or expression changes will provide information of clinical value^{3,43}. If (which is not clear) the genetic pathway adopted by a given tumour is influenced either by genetic background or by environmental exposures, the 'molecular phenotype' may also define groups of tumours that aetiologically are more homogeneous, which would be valuable information in studies of genetic or environmental predisposition. Finally, adoption of a particular pathway of progression may constrain the possibilities for evolution of the cancer in the future. Clinical experience suggests that there are categories of pre-invasive change in, for example, prostate or breast epithelium which are, at the stage they are recognized, already largely determined in their potential for future malignancy. This implies that chance subsequent events in the evolution of these lesions cannot lead to a more malignant phenotype. If so, it will be important to find out whether molecular phenotypes can predict future malignant potential more accurately than current histological methods and, if they can, to use this information to judge strategies for intervention. A topical example is provided by the controversies surrounding radical treatment of early prostatic cancer detected by screening⁴⁴.

Genetic events outside the cancer pathway

So far, our focus has been on the developing cancer cell, and on the pathway genetic events and the deficiencies in DNA repair and genomic stability which may drive them. Productive though this focus has been and will continue to be, it provides only part of the picture. It is likely that genetic variation at other sites, both inside and outside the cancer cell, may substantially affect cancer development. This is illustrated by the following brief examples.

Gene–environment interaction

Genetic variation acting either within or outside the cancer cell may determine the outcome of interaction with exogenous carcinogens. A clear example is provided by the greater risk of cutaneous melanoma as a result of sun exposure in individuals with a fair skin, or who have many naevi (a phenotype which is genetically determined). Polymorphisms at the interleukin-1 locus, which are associated with increased production of interleukin-1 β , are associated with both an increased risk of hypochlorhydria induced by the gastric pathogen *Helicobacter pylori*, and gastric cancer⁴⁵. Analogous interactions are to be expected between chemical exposures and genetic variations in metabolic pathways, although well-attested examples are still rather few⁵ (see ref. 5 and the article in this issue by Peto, pages 390–395). Such variation may in principle account for substantial differences in cancer susceptibility within the population, and knowledge of gene–environment interaction may indicate strategies for prevention in those at risk. Information about relevant genetic variation may also help in the design of epidemiological studies: categorization

of subpopulations in terms of genetic risk may reduce heterogeneity and so increase power to detect causative exposures. Finally, tissue-specific patterns of gene expression may indicate which genes, and therefore which exposures, are likely to be relevant⁴⁶.

Local factors affecting the developing cancer cell

Wounding and chronic inflammation have long been known to be associated with cancer. Their effects may be mediated either through increased mitogenesis, which may be associated with increased mutation⁴⁷, or through paracrine effects, for example from inflammatory cells. Thus, production of the matrix metalloproteinase MMP9 by inflammatory cells has been implicated in the development of squamous cell carcinomas in an HPV-16 transgenic model, and various inflammatory cytokines have been shown to affect p53 transcriptional regulation and apoptosis in epithelial cells (reviewed in ref. 48). Such processes presumably underlie the increased cancer risk in diseases such as ulcerative colitis and hereditary pancreatitis⁴⁹, which have an inherited component. It is also likely that there will be genetically determined variation in the wounding and inflammatory responses themselves, which will affect cancer initiation and progression.

There is accumulating evidence for an important role of paracrine interactions between epithelium and stroma in epithelial carcinogenesis⁷. Reciprocal 'conditioning' between cancer and adjacent stromal cells has been shown in tissue recombination experiments⁵⁰. Irradiation of mammary gland stroma can promote the expression of tumorigenic potential by unirradiated epithelial cells⁵¹. Several studies provide evidence for a role of matrix metalloproteinases in the early as well as late stages of cancer development^{7,52}. In general, transgenic mice that overexpress MMPs develop more cancers in response to oncogenic stimuli, whereas those that lack different MMPs or overexpress inhibitors develop fewer (but more malignant) cancers⁵³. Although no data are currently available, it seems plausible that there will be polymorphic variation in MMP activity in human tissues, and that this may affect both the development of cancer and the behaviour of the cancers that result. Similar genetically determined variation may be expected in processes later in cancer development; such as angiogenic responses (see article in this issue by Liotta and Kohn, pages 375–379).

Systemic factors

Variations in circulating levels of hormones or growth factors show significant association with cancer risk. In one population-based study of oestradiol levels in post-menopausal women, there was an almost fivefold difference in risk of breast cancer between the upper and lower tertiles of circulating oestradiol level⁵⁴. High levels of oestrogen might have carcinogenic effects either through direct stimulation of growth or as a by-product of mutagenic metabolites. Similar effects have been reported for several common cancers in

relation to the IGF family⁹, and there is some evidence that a significant proportion of the variance in circulating IGF-1 levels is genetically determined⁵⁵. Such genetic variation is a further plausible mechanism for a significant component of individual cancer susceptibility.

Inherited predisposition

The cardinal feature by which inherited predisposition is recognized clinically is family history. Cancer is common, so some families will contain several cases by chance. There is a spectrum of probability that a given family history reflects inherited predisposition from near-certainty of strong predisposition in the rare inherited cancer syndromes, to the possibility of weak effects in familial clusters (Table 1). Paradoxically, the largest category of inherited predisposition, in terms of expected fraction of cancer incidence, is the one with the weakest genetic effects — ‘predisposition without evident family clustering’^{56,57}. The combined contribution to overall breast cancer incidence of strongly predisposing mutations in *BRCA1* and *BRCA2*, which confer individual risks of around 60% by age 70, is less than 5%. By contrast, a predisposing allele with a relative risk of 2 and frequency of 20% could account for up to 20% of breast cancer incidence.

Strong predisposition

The human inherited cancer syndromes and their transgenic mouse counterparts have been reviewed extensively^{58,59}. In the cases described so far, strong predisposition to cancer results either through inheritance of one of the events on the cancer ‘pathway’, or through effects on DNA repair or genome stability. Studies of the mechanisms of predisposition in these syndromes have led to substantial insights into cancer biology. Genetic testing for risk is now part of the standard of clinical care for families, although its value may be controversial when the practical benefits of the actions open to someone at risk are not clear⁶⁰.

Two features of these syndromes merit brief comment, because if we could explain them, we would know more about the development of cancer. They are tissue specificity and variability of expression. All inherited predisposition to cancer seems to show a considerable degree of tissue specificity, even in the case of predisposition by defective DNA repair. In most cases, there is no obvious lineage or physiological explanation for the patterns and the mechanisms are unknown. There may also be considerable variation in the age at onset of cancer and in the specific types of cancer that predominate not only within a given syndrome, but also within a single family. Some of this variation is due to different germline alleles of the main predisposing gene (for example, in Von Hippel Lindau disease⁶¹, familial adenomatous polyposis⁶² and multiple endocrine neoplasia type 2⁶³) and some is environmental or chance. But much of the within-family variation is probably attributable to the effects of genetic modifiers. This has been clearly shown in a number of mouse cancer models⁶⁴, and by the demonstration that concordance of phenotype in neurofibromatosis type 1 is greatest in monozygotic twins and decays with increasing distance of relationship¹⁰. Many of these modifiers are likely to overlap with the low-penetrance predisposing genes described in the next section. One practical implication of modifier effects is that the quoting of risks for individuals who carry genes such as *BRCA1* is an uncertain business. This is relevant to insurance, where the uncertainties are perhaps not sufficiently recognized. Inappropriately high-risk figures may be used, which derive from reports of the extreme set of families that are usually the first to be studied. A more speculative implication is that if we knew the mechanisms of modification, we might exploit this knowledge for treatment¹¹ or prevention.

Weak predisposition

Weak predisposition to cancer may in principle result from weak alleles of the pathway or caretaker genes described in the last section, or from genetic variation at the other sites indicated in Fig. 1. The study of weak predisposition is of interest both for its possible public-

health implications⁵⁶ and because just as the study of inherited cancer syndromes identified ‘pathway’ genes, so weak predisposition may point to a wider range of processes that are relevant to cancer development, and to interactions between them. The search for these genes is just beginning and as yet there are few data. The principles can be illustrated from studies of breast cancer.

In breast cancer, the risk to close relatives of a case, averaged across all ages, is about twofold. Most of this familial risk is probably genetic in origin (see article in this issue by Peto, pages 390–395). The risk is about the same for the mother, sisters or daughters of a case, suggesting dominant rather than recessive effects. Large population-based studies indicate that only 15–20% of overall familial risk is attributable to mutations in *BRCA1* and *BRCA2*⁶⁵. The possibilities for the remaining 80% are some combination of a small number of moderately strong genes, and a larger number (possibly a hundred or more) of weaker genes (Box 1). If moderately strong genes exist, it should in principle be possible to identify them by linkage in families. The weaker genes will not, on the whole, result in multiple case families and so must be sought by a different approach: a comparison of the frequency of candidate genetic variants between a large series of cancer cases and controls (an ‘association study’). The candidate genes might lie anywhere in the scheme outlined in Fig. 1. Of the first 40 or so candidates tested for association with breast cancer, a few show evidence of weak effects, most of which require independent confirmation. They include genes encoding steroid hormone receptors and paracrine growth factors, and genes involved in metabolism of exogenous chemicals, and in DNA repair⁶⁶. The variant alleles are associated with risks of around 1.5-fold and are predicted to account for only a few per cent of breast cancer incidence. Collectively they account for only a very small fraction of the familial risk. Almost certainly there are many more genes to be identified, which together will account for a much higher fraction of cancer incidence than the genes in the inherited cancer syndromes.

The identification of these genes will be greatly accelerated by the data from the Human Genome Project⁶⁷. The search relies on cataloguing the DNA sequence variation within the population, and showing (currently on a ‘candidate’ gene-by-gene basis) that particular variants are significantly associated either with disease susceptibility or with some other aspect of disease phenotype such as treatment response or survival⁶⁸. The most readily assayed form of genomic variation is the single nucleotide polymorphism or ‘SNP’: of the order of one million SNPs have been identified and are available from genomic databases⁶⁹. Comparable data from the mouse genome project will support similar studies in mice. Here, the availability of cancer models, and the possibilities of experimental manipulation on a defined genetic background, allow an empirical search for genetic modifiers and low-penetrance genes on a genome-wide basis, which may provide valuable candidates to test in human populations^{11,70}. Lessons from the much longer history of quantitative genetic analysis in lower organisms are also likely to be valuable⁷¹. There are, of course, many problems still to be addressed (for review, see ref. 68), but possibly the most pressing is the lack of sufficiently large and well-documented human case-control sets to analyse. This, rather than the genetic or statistical technology, is currently the limiting factor. In general, funding agencies have in the past been curiously unwilling to face up to this; now when they may be changing, there is the potential threat from ‘the new ethics’ discussed in the article by Peto, pages 390–395, which may put further costs and difficulties in the way. Despite this, it seems certain that the next decade will see significant advances in understanding the polygenic basis of many diseases, including cancer.

The future

Some have hailed the approaching era of the polygenic basis of disease as a new dawn⁷²; others are sceptical⁷³. The sceptics argue, in relation to cancer predisposition, that the genes are weak in comparison to lifestyle and environmental causes or risk, and it will be difficult to use this type of genetic information to practical effect. The

numbers relating to avoidable cancer risks presented in the article by Peto seem to support this. However, as also discussed by Peto, the picture may be different if the aggregate effect of several genes and other non-genetic predisposing factors can define a spectrum of risk across the population which is sufficiently wide. In that case, these factors might be used to construct 'risk profiles' that would identify either small groups of people at high risk who account for a substantial fraction of cancer incidence, or large groups who are at very low risk (and who can therefore be discouraged from taking up costly and perhaps risky interventions). Our modelling of the distribution of breast cancer risk in a UK population (Pharoah *et al.*, unpublished data) predicts that there may be as much as a 40-fold difference in relative risk between the highest and lowest quintiles of the distribution that could be defined by a genotypic profile. As genes are identified, the predictive power of the available profiles can be tested in the large population cohorts that are being followed for cancer incidence. The goal of genotypic profiling is probably distant, because it may require that a majority of the tens or even hundreds of predisposing alleles be identified; and if it does become possible, there will be social and ethical issues to address. Nevertheless, it seems an attainable goal. □

1. Nowell, P. C. The clonal evolution of tumour cell populations. *Science* **194**, 23–28 (1974).
2. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
3. Liotta, L. & Petricoin, E. Molecular profiling of human cancers. *Nature Rev. Genet.* **1**, 48–56 (2000).
4. Baylin, S. B. & Herman, J. G. DNA methylation in tumorigenesis. *Trends Genet.* **16**, 168–174 (2000).
5. Rothman, N. *et al.* The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. *Biochem. Biophys. Acta* **1471**, C1–C10 (2001).
6. Olden, K. & Wilson, S. Environmental health and genomics: visions and implications. *Nature Rev. Genet.* **1**, 149–153 (2000).
7. Tlsty, T. D. & Hein, P. W. Know thy neighbor: stromal cells can contribute oncogenic signals. *Curr. Opin. Genet. Dev.* **11**, 54–59 (1998).
8. Risch, A. Hormonal etiology of epithelial ovarian cancer, with a hypothesis concerning the role of androgens and progesterone. *J. Natl Cancer Inst.* **90**, 1774–1768 (1998).
9. Yu, H. & Rohan, T. Role of the insulin-like growth factor family in cancer development and progression. *J. Natl Cancer Inst.* **92**, 1472–1489 (2000).
10. Easton, D. F., Ponder, M. A., Huson, S. M. & Ponder, B. A. J. An analysis of variation in expression of neurofibromatosis (NF) type 1 (NF1): evidence for modifying genes. *Am. J. Hum. Genet.* **53**, 305–313 (1993).
11. Nadeau, J. H. Modifier genes in mice and humans. *Nature Rev. Genet.* **2**, 165–174 (2001).
12. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA* **68**, 820–823 (1971).
13. Harris, H. The analysis of malignancy by cell fusion: the position in 1988. *Cancer Res.* **48**, 3302–3306 (1988).
14. Park, M. in *The Genetic Basis of Human Cancer* (eds Vogelstein, B. & Kinzler, K. W.) 205–228 (McGraw Hill, New York, 1998).
15. Bishop, J. M. Enemies within: the genesis of retrovirus oncogenes. *Cell* **23**, 5–6 (1981).
16. Parada, L. P., Tabin, C. J., Shih, C. & Weinberg, R. A. Human EJ bladder carcinoma oncogene is a homologue of Harvey Sarcoma virus ras gene. *Nature* **297**, 474–477 (1982).
17. Rabbitts, T. H. Chromosomal translocations in human cancer. *Nature* **372**, 143–149 (1994).
18. Brodeur, G. M. & Hogarty, M. D. in *The Genetic Basis of Human Cancer* (eds Vogelstein, B. & Kinzler, K. W.) 161–172 (McGraw Hill, New York, 1998).
19. Mitelman, F., Mertens, F. & Johansson, B. A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet.* **13**, 417–474 (1997).
20. Gray, J. W. & Collins, C. Genome changes and gene expression in human solid tumours. *Carcinogenesis* **21**, 443–452 (2000).
21. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–957 (2001).
22. DePinho, R. A. The age of cancer. *Nature* **408**, 248–254 (2000).
23. Comings, D. A general theory of carcinogenesis. *Proc. Natl Cancer Inst.* **70**, 3324–3328 (1973).
24. Cavenee, W. K. *et al.* Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature* **305**, 779–781 (1983).
25. Weinberg, R. A. Tumor suppressor genes. *Science* **254**, 1138–1146 (1991).
26. Fearon, E. R. in: *The Genetic Basis of Human Cancer* (eds Vogelstein, B. & Kinzler, K. W.) 229–236 (McGraw Hill, New York, 1998).
27. Haber, D. & Harlow, E. Tumor suppressor genes: evolving definitions in the genomic age. *Nature Genet.* **16**, 320–322 (1997).
28. Kinzler, K. W. & Vogelstein, B. Gatekeepers and caretakers. *Nature* **386**, 761–763 (1997).
29. Osborne, R. J. & Hamshire, M. G. A genome-wide map showing common regions of loss of heterozygosity/allelic imbalance in breast cancer. *Cancer Res.* **60**, 3706–3712 (2000).
30. Faro, M. L. *et al.* The murine gene p27^{Kip1} is haploinsufficient for tumour suppression. *Nature* **396**, 177–180 (1998).
31. Rainier, S. *et al.* Relaxation of imprinted genes in human cancer. *Nature* **362**, 747–749 (1993).
32. Cui, H. *et al.* Loss of imprinting in normal tissue of colorectal cancer patients with microsatellite instability. *Nature Med.* **4**, 1276–1280 (1998).
33. Herman, J. G. *et al.* Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc. Natl Cancer Inst.* **95**, 6870–6875 (1998).
34. Chen, R. Z., Pettersson, U., Beard, C., Jackson-Crusher, L. & Jaenisch, R. DNA hypomethylation leads to elevated mutation rates. *Nature* **395**, 89–93 (1998).
35. Jackson-Crusher, L. *et al.* Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nature Genet.* **27**, 31–39 (2001).
36. Cormier, R. J. & Dove, W. F. Dnmt^{+/−} reduces the net growth rate and multiplicity of intestinal adenomas in C57BL/6-multiple intestinal neoplasia (Min)/+ mice. *Cancer Res.* **60**, 3965–3970 (2000).
37. Issa, J.-P. J. *et al.* Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nature Genet.* **7**, 536–540 (1994).
38. Feinberg, A. P. Methylation meets genomics. *Nature Genet.* **27**, 9–10 (2001).
39. Sapienza, C. Genome imprinting, cellular mosaicism and carcinogenesis. *Mol. Carcinogenesis* **3**, 118–124 (1990).
40. Engler, P. *et al.* A strain-specific modifier on mouse chromosome 4 controls the methylation of independent transgene loci. *Cell* **65**, 939–947 (1991).
41. Silver, A. J. & White, R. Inheritance of allelic blueprints for methylation patterns. *Cell* **54**, 145–152 (1988).
42. Renan, M. J. How many mutations are required for tumorigenesis? Implications from human cancer data. *Mol. Carcinogenesis* **7**, 139–146 (1993).
43. Suzuki, S. *et al.* An approach to analysis of large scale correlations between genome changes and clinical endpoints in ovarian cancer. *Cancer Res.* **60**, 5382–5385 (2000).
44. Neal, D. E. & Donovan, J. L. Prostate cancer: to screen or not to screen? *Lancet Oncol.* **1**, 17–24 (2000).
45. El-Omar, E. M. *et al.* Interleukin-1 polymorphisms associated with increased risk of gastric cancer. *Nature* **404**, 398–402 (2000).
46. Williams, J. A. Single nucleotide polymorphisms, metabolic activation and environmental carcinogenesis: why molecular epidemiologists should think about enzyme expression. *Carcinogenesis* **22**, 209–214 (2001).
47. Ames, B. N. & Gold, L. S. Too many rodent carcinogens: mitogenesis increases mutagenesis. *Science* **249**, 970–971 (1990).
48. Cordon-Cardo, C. & Prives, C. At the crossroads of inflammation and tumorigenesis. *J. Exp. Med.* **190**, 1367–1370 (1999).
49. Lowenfils, A. B. *et al.* Hereditary pancreatitis and the risk of pancreatic cancer. *J. Natl Cancer Inst.* **89**, 442–446 (1997).
50. Olumi, A. F. *et al.* Carcinoma-associated fibroblasts direct tumor progression of initiated human prostatic epithelium. *Cancer Res.* **59**, 5002–5011 (1999).
51. Barcellos-Hoff, M. H. & Ravani, S. A. Irradiated mammary gland stroma promotes the expression of tumorigenic potential by unirradiated epithelial cells. *Cancer Res.* **60**, 1254–1260 (2000).
52. Sternlicht, M. D. *et al.* The stromal protease MMP3/Stromelysin-1 promotes mammary carcinogenesis. *Cell* **96**, 137–146 (1999).
53. Coussens, L. M., Tinnh, C. C., Hanahan, D. & Werb, Z. MMP-9 supplied by bone marrow-derived cells contributes to skin carcinogenesis. *Cell* **103**, 481–490 (2000).
54. Thomas, M. V. *et al.* A prospective study of endogenous serum hormone concentrations in post-menopausal women on the island of Guernsey. *Br. J. Cancer* **76**, 401–405 (1997).
55. Jernstrom, H. *et al.* Genetic and non-genetic factors associated with variation of plasma levels of insulin-like growth factor-1 and insulin-like growth factor binding protein-3 in healthy pre-menopausal women. *Cancer Epidemiol. Biomarkers Prev.* (in the press).
56. Peto, J. in *Cancer Incidence in Defined Populations* Banbury Report 4 (eds Cairns, J., Lyon, J. L. & Skolnick, M.) (Cold Spring Harbor, New York, 1980).
57. Ponder, B. A. J. Inherited predisposition to cancer. *Trends Genet.* **6**, 213–218 (1990).
58. Vogelstein, B. & Kinzler, K. W. (eds) *The Genetic Basis of Human Cancer* (McGraw Hill, New York, 1998).
59. Klausner, R. D. Studying cancer in the mouse. *Oncogene* **18**, 5249–5252 (1999).
60. Ponder, B. A. J. Genetic testing for cancer risk. *Science* **278**, 1050–1054 (1997).
61. Chen, F. *et al.* Germline mutations in the Von Hippel Lindau disease tumour suppressor gene: correlation with phenotype. *Hum. Mutat.* **5**, 66–75 (1995).
62. Kinzler, K. W. & Vogelstein, B. in *The Genetic Basis of Human Cancer* (eds Vogelstein, B. & Kinzler, K. W.) 565–587 (McGraw Hill, New York, 1998).
63. Eng, C. & The International RET Mutation Consortium. The relationship between specific RET protooncogene mutations and disease phenotype in multiple endocrine neoplasia type 2. *J. Am. Med. Assoc.* **276**, 1575–1579 (1996).
64. Halberg, R. B. *et al.* Tumorigenesis in the multiple intestinal neoplasia mouse: redundancy of negative regulations and specificity of modifiers. *Proc. Natl Acad. Sci. USA* **97**, 3461–3466 (2000).
65. The Anglian Breast Cancer Study Group. Prevalence of BRCA1 and BRCA2 mutations in a large population based series of breast cancer cases. *Br. J. Cancer* **83**, 1301–1308 (2000).
66. Dunning, A. M. *et al.* A systematic review of genetic polymorphisms and breast cancer risk. *Cancer Epidemiol. Biomarkers Prev.* **8**, 843–854 (1999).
67. Chakravarti, A. ... to a future of genetic medicine. *Nature* **409**, 822–823 (2001).
68. Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99 (2001).
69. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
70. Balmain, A. & Nagase, H. Cancer resistance genes in mice: models for the study of tumour modifiers. *Trends Genet.* **14**, 139–144 (1998).
71. Mackay, T. F. C. Quantitative trait loci in *Drosophila*. *Nature Rev. Genet.* **2**, 11–20 (2001).
72. Bell, J. The new genetics in clinical practice. *Br. Med. J.* **316**, 618–620 (1998).
73. Holtzman, N. A. & Marteau, T. M. Will genetics revolutionize medicine? *New Engl. J. Med.* **343**, 141–144 (2000).

Acknowledgements

I apologize that space has restricted the topics and the references that could be included. I thank D. Easton for advice on the figure in Box 1, and several colleagues for helpful suggestions. B.A.J.P. is a Gibb Fellow of The Cancer Research Campaign (CRC).

Proliferation, cell cycle and apoptosis in cancer

Gerard I. Evan* & Karen H. Vousden†

*UCSF Cancer Center, 2340 Sutter Street, San Francisco, California 94143-0875, USA

†NCI at Frederick, Building 560, West 7th Street, Frederick, Maryland 20842, USA

Beneath the complexity and idiopathy of every cancer lies a limited number of 'mission critical' events that have propelled the tumour cell and its progeny into uncontrolled expansion and invasion. One of these is deregulated cell proliferation, which, together with the obligate compensatory suppression of apoptosis needed to support it, provides a minimal 'platform' necessary to support further neoplastic progression. Adroit targeting of these critical events should have potent and specific therapeutic consequences.

Since its inception, the study of the molecular basis of cancer has carried with it the promise of more refined, more effective cancer therapies. It has generally been assumed that because cancers are derived from numerous tissues with multiple aetiologies, and as tumour progression carries with it a bewildering and seemingly endless combination of genetic and epigenetic alterations giving rise to a hugely disparate series of diseases, cures for cancer must be as diverse as the diseases themselves. The mantra from the cancer research community has been that cancer is not a single disease for which there will be a single cure, and the task of developing therapies suitable for treatment of the full gamut of cancers is depicted as Herculean and almost impossible.

In this review, we entertain the idea that these assertions are unnecessarily pessimistic. Although cancers are indeed extremely diverse and heterogeneous, we suggest that underlying this variability lies a relatively small number of 'mission critical' events whose convergence is required for the development of any and all cancers. The focus of this perspective is on two of these: the lesions that power the relentless proliferation of tumour cells, and the compensatory mutations that arise to ensure their survival. Although neoplasia involves many other processes that also present targets for cancer therapy¹, in almost all instances, deregulated cell proliferation and suppressed cell death together provide the underlying platform for neoplastic progression. The challenge before the research community is to identify and understand the molecular anatomy of such pivotal steps in tumour progression and to develop therapies that directly attack these points of convergence.

Evolution of cancers

Cancers are diseases in which unremitting clonal expansion of somatic cells kills by invading, subverting and eroding normal tissues. Driving cancer development are stochastic somatic cell mutations in genes that govern and regulate the diverse aspects of metazoan growth control. The processes governing the genesis and progression of cancers are evolutionary ones in which natural selection acts upon the inherent or acquired diversity of various somatic clones, fostering the outgrowth of those with some form of propagative advantage. Metazoans must restrain this tendency of individual somatic cells to establish their own autonomous colonies, yet at the same time sanction sufficient somatic cell proliferation to build and maintain the whole organism. The

solution adopted by most animals is simple: adults are small, short-lived and disposable egg dispersers, constructed almost exclusively of post-mitotic cells whose irreversible loss of proliferative capacity effectively curtails any opportunity for mutation and somatic evolution.

Unfortunately, long-lived organisms such as vertebrates need substantial and continuous cell proliferation throughout their extended lives, both for development and long-term maintenance and repair. In teleological terms, the evolutionary imperative of vertebrates has been to find a way to allow cell proliferation when needed, while at the same time efficiently suppressing the genesis of mutated cells leading to deregulated growth. When such measures fail, cancer is the inevitable consequence.

Awareness of the evolutionary nature of cancer offers a number of important insights into the malignant process. First, and perhaps most striking, is the rarity of the cancer cell. With an estimated mutation rate of some 1 in 2×10^7 per gene cell division², some 10^{14} target cells in the average human, and an abundant repertoire of genes regulating all aspects of cell expansion, it is remarkable that cancers arise in only 1 in 3 lifetimes. This is even more striking when one considers that oncogenic mutations, by their nature, foster clonal expansion of the affected cell, so propagating the initial mutation and thereby increasing the number of target cells available for (and hence the probability of) further oncogenic mutation. The rarity of cancer highlights the efficacy of potent anti-tumorigenic mechanisms presiding over somatic cells. Cancers prevail only when these mechanisms have failed³.

Second, cancers 'progress' for the same reason organisms seem to — we see only the successes, not the failures. This distorts our statistical view of cancer progression. No matter how rare the genesis and evolution of a cancer cell or how effective the anti-cancer therapy administered, our perception is only of the rare surviving clones that beat all the odds and appear as clinical disease. Our inability to discern the mechanisms that thwart the vast majority of inchoate tumours deprives us of great insight into how these mechanisms break down in cancer and, correspondingly, how we might best reactivate them.

Third, evolutionary trajectories of cancers are shaped by the selective pressures they encounter. Tumours evolve within differing somatic environments, each of which imposes its own unique constraints. For example, shedding epithelia such as gut or skin 'defend' themselves against the emergence of sizeable mutant clones by condemning all

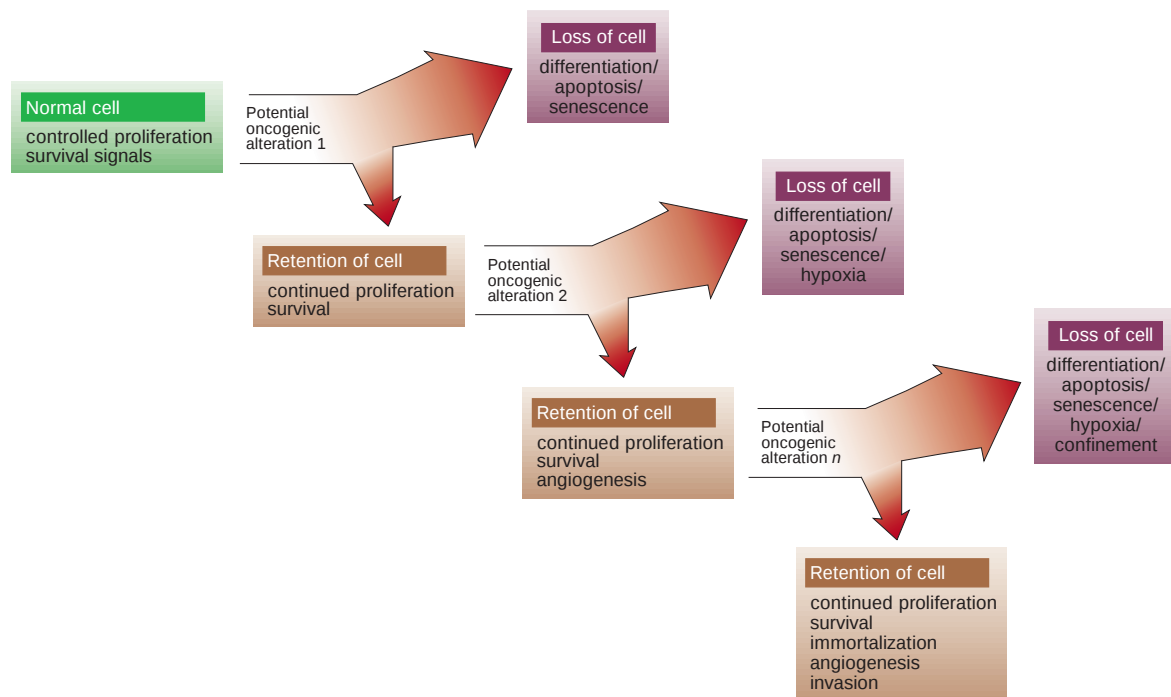


Figure 1 Evolution of cancer is more complex than the straightforward linear accumulation of oncogenic mutations. Potentially oncogenic proliferative signals are coupled to a variety of growth-inhibitory processes, such as the induction of apoptosis, differentiation or senescence, each of which restricts subsequent clonal expansion and neoplastic evolution. Tumour progression occurs only in the very rare instances where these growth-inhibitory mechanisms are thwarted by compensatory mutations.

progeny cells to terminal differentiation and death. Derailing of this differentiating conveyor belt is an important part of gastrointestinal and skin cancer, but is clearly irrelevant to the process of carcinogenesis in a tissue such as liver.

Fourth, evolution is an ongoing process. As a neoplasm progresses, expands and spreads, it confronts shifting selective pressures. The heterogeneity and diversity seen in cancers are vestiges of a dynamic and stochastic evolutionary force that varies with differing somatic environments.

The commonality of cancers

Tumours are diverse and heterogeneous, but all share the ability to proliferate beyond the constraints limiting growth in normal tissue. Aberrations in the regulation of a restricted number of key pathways that control cell proliferation and cell survival are mandatory for establishment of all tumours. Deregulated cell proliferation together with suppressed apoptosis constitute the minimal common platform upon which all neoplastic evolution occurs. The critical issue is to identify how tumour cells differ from normal cells and how those differences can be exploited therapeutically.

Limits to clonal autonomy of metazoan cells

The restriction of clonal autonomy that is essential in vertebrate biology is implemented by tiers of mechanisms, each one of which must be somehow evaded or negated for cancers to arise (Fig. 1).

Normal somatic cells are totally dependent for their proliferation upon receipt of appropriate mitogenic signals. Mitogens act as obligate social cues that constrain cells to proliferate only in the appropriate social context. Furthermore, cells become committed to entry of the cell cycle only towards the end of G1, a retinoblastoma (pRB)-regulated transition point which most cell types reach only after hours or days of sustained mitogen exposure⁴. Thus, cells will respond only to proliferative impetuses of some tenacity. In some cases, sustained mitogenic signalling can only occur within a specific

somatic context. For example, the transient and mitogenically inadequate induction of cyclin D1, induced by mitogen activation of receptor tyrosine kinase (RTK) signalling, is transmuted into a persistent and mitogenically productive response upon co-stimulation of integrins via attachment to the extracellular matrix (ECM)⁵.

Superimposed upon the requirement for positive growth signals lies a web of growth inhibitory factors that serve to gate the proliferative response to mitogens, and which has to be overcome for cell-cycle entry¹. Examples of such factors are transforming growth factor- β ⁶ and the interferons⁷. These pleiotropic signalling molecules exert potent anti-proliferative effects, in part by suppressing phosphorylation of pRB, through their inhibitory effects on cyclin-dependent kinases (CDKs) and induction of various CDK inhibitors, and also by their suppression of c-Myc.

The inverse coupling of differentiation to proliferation is another hardwired restraint to somatic cell autonomy, as proliferative potential of somatic cells is counterbalanced by an innate predisposition of progeny cells to engage pathways of terminal differentiation⁸. Moreover, unfettered proliferative potential is restricted to a small number of slowly replicating stem cells. These typically undergo infrequent asymmetric divisions, generating one daughter that replaces the original, while the other enters a transit amplifying population resulting in irreversible commitment to a terminal differentiation programme. By confining most cell expansion to cells already committed to ultimate genetic or physical death, stem cells allow provision of sufficient cells to maintain and replace tissues, while restricting the number of cell divisions (and hence exposure to mutagenic risk) in those somatic cells with significant proliferative potential^{9–11}.

Somatic cells that evolve the capacity for proliferative autonomy still face major obstacles to their continued expansion. Metazoan somatic cells are obligatorily dependent for their survival upon the continuous availability of trophic factors, which are often in limiting supply and spatially restricted^{12,13}. Consequently, deregulated cell expansion results

in exhaustion of local survival factors and the triggering of apoptosis. Furthermore, many rapidly proliferating epithelial tissues have evolved architectures that ensure the eventual death of progeny cells as they are forced to migrate outwards to be shed from the surface. Should rare clones then succeed in evading both growth control and death, they then encounter the ultimate proliferative backstop. Repeated divisions erode their telomeres, ultimately triggering irreversible arrest or, more likely, apoptosis¹⁴. Finally, to form a tumour the errant clone must make its way in the outside world of somatic tissues. Substantial evidence indicates that development of macroscopic metastatic cancers requires the capacity to erode and subvert normal tissues and commandeer a nurturing vasculature from pre-existing blood vessels in adjacent normal tissues (see article in this issue by Liotta and Kohn, pages 375–379).

Cancer as a disease of deregulated cell proliferation

Each of the pathways that constrains the proliferative response in normal cells is perturbed in most cancers. One class of mutations required for tumour development acts by short circuiting the normally obligate requirement of somatic cells for external mitogenic signals¹⁵. Such mutations may involve autocrine production of a normally limiting mitogen, activating mutations of the mitogen RTKs or G-protein signal transducers such as Ras, or mutations affecting one of the many intermediary signal transducing molecules that convey mitogenic information to its intracellular targets (see review in this issue by Blume-Jensen and Hunter, pages 355–365). A second class of growth-deregulating mutations comprises those that target the principal late-G1 cell-cycle checkpoint regulated by pRB¹⁶. Defects in this pathway, which may be universal in human cancers, include deletion of the *RB* gene itself and deregulation of the CDKs that phosphorylate and functionally inactivate pRB, either through direct over-activation of CDKs or through genetic loss of their inhibitors¹⁷. Another frequent proliferative lesion that has the effect of deregulating the cell cycle is uncontrolled expression of Myc¹⁸. Myc expression is tightly controlled by mitogen availability in normal cells, but it is usually expressed in a deregulated or elevated manner in tumour cells. Myc seems to be a strategic controller of cell proliferation that acts pleiotropically to coordinate both cell growth^{19–21} and concomitant progression through the cell cycle^{22,23}.

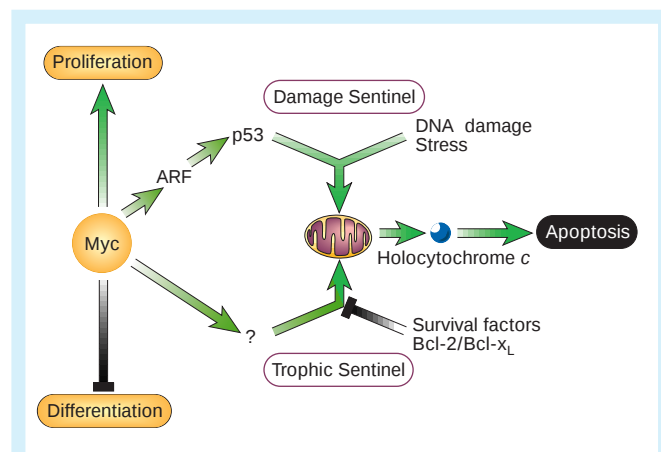


Figure 2 Activation of growth-deregulating lesions triggers 'sentinel' functions that guard the cell against acquiring mutations or propagating into an inappropriate somatic compartment. The more powerful and persistent the growth signal, the more potent and persistent the sentinel function. In this example, the oncoprotein Myc is shown activating a p53 damage sentinel through the ARF/MDM-2 pathway, thereby sensitizing the cell to any DNA damage. Myc also promotes release of holo-cytochrome c from the mitochondrion into the cytosol where it triggers apoptosis. Release of holo-cytochrome c is inhibited by paracrine 'survival' signals that are typically restricted both in supply and location. Clonal outgrowth driven by relentless Myc expression outstrips survival factor availability, triggering the 'trophic sentinel' to kill the cell.

The presence in individual tumours of multiple mutations that affect each of the pathways discussed above suggests that each pathway contributes a discrete type of proliferative function to the neoplastic phenotype. But precisely what such functions are and how and why they interact, remains unknown. Moreover, in certain circumstances single types of proliferative lesion seem sufficient to drive cell proliferation. For example, mere deregulation of c-Myc is, at least in the mouse, alone sufficient to induce and maintain proliferation of multiple somatic cell types *in vitro* and *in vivo*^{24,25}.

In addition to driving aberrant cell division, mutations in the various proliferative control pathways have a profound impact on other cell functions. For example, many of the proliferative lesions in tumour cells also contribute to the inhibition of differentiation, thereby preventing the elimination of progeny cells from the proliferative compartment of many types of tissue. pRB, for example, is essential in differentiation of several tissue types through interactions with factors such as the helix–loop–helix proteins MyoD²⁶ and Id2 (ref. 27). Loss or inhibition of pRB function prevents normal differentiation, a contribution to tumour development distinct from the direct deregulation of cell-cycle progression. Deregulated Myc expression also inhibits differentiation, in part by activation of Id2 expression²⁷.

Cancer as a disease of deregulated survival

Survival of all somatic cells requires the continuous input of survival and trophic signals to suppress apoptosis. The central engines of apoptosis are the caspases, cascades of cysteine aspartyl proteases that implement cell death by cleaving a variety of intracellular substrates that trigger cell dissolution. Caspases are synthesized as latent zymogens that are activated by proteolytic cleavage: typically through the action of upstream apical caspases. One such pathway is mediated by transmembrane death receptors of the CD95 (Apo-1 or Fas)/TRAIL/tumour-necrosis factor (TNF) receptor 1 family, whose ligation triggers recruitment and assembly of multiprotein complexes that activate apical caspase 8 (ref. 28). The other principal death-signalling pathway involves the mitochondrion, which acts as an integrating sensor of multiple death insults by releasing cytochrome c into the cytosol where it triggers caspase activation. The mitochondrial pathway is thought to be the principal target of survival signalling pathways, which act by stabilizing mitochondrial function and integrity and suppressing release of cytochrome c²⁹. Once cytochrome c has been released from the mitochondrion, it orchestrates assembly of an intracellular apoptosome complex that recruits apical caspase 9 via the adaptor protein Apaf-1 (ref. 30).

Viability of normal somatic cells requires survival signals that are idiosyncratic to each cell type; signals include soluble factors or direct physical interactions with neighbouring cells or ECM. Because such signals are available typically only within discrete somatic environments, metazoan somatic cells are in effect 'trapped' within specialized trophic microenvironments within the body, dying should they wander or become misplaced. Epithelial cells offer a particularly dramatic example of such somatic entrapment. Detachment from their neighbours or basal stroma triggers a spontaneous apoptotic suicide termed anoikis. In part, anoikis occurs because detachment deprives the cell of necessary integrin and cadherin-mediated survival signals. However, it has recently been shown that disturbances to the intracellular cytoskeleton induced by detachment can directly trigger apoptosis through release of pro-apoptotic BH3 proteins such as Bmf, which is normally kept inactive through binding to the actin-based motor complex (D. Huang, H. Puthalakath and A. Strasser, personal communication). Another BH3 protein, Bim, is bound to the LC8 cytoplasmic dynein light chain, which sequesters it to the microtubule-associated dynein motor complex, but is released in response to multiple apoptotic stimuli³¹.

With such potent mechanisms in existence to obliterate displaced cells, it is no surprise that suppression of apoptosis is high on the list of acquired attributes in cancer cells. Known mutations in survival signalling pathways found in tumours include deregulated expression of the survival factors insulin-like growth factor (IGF)-I and

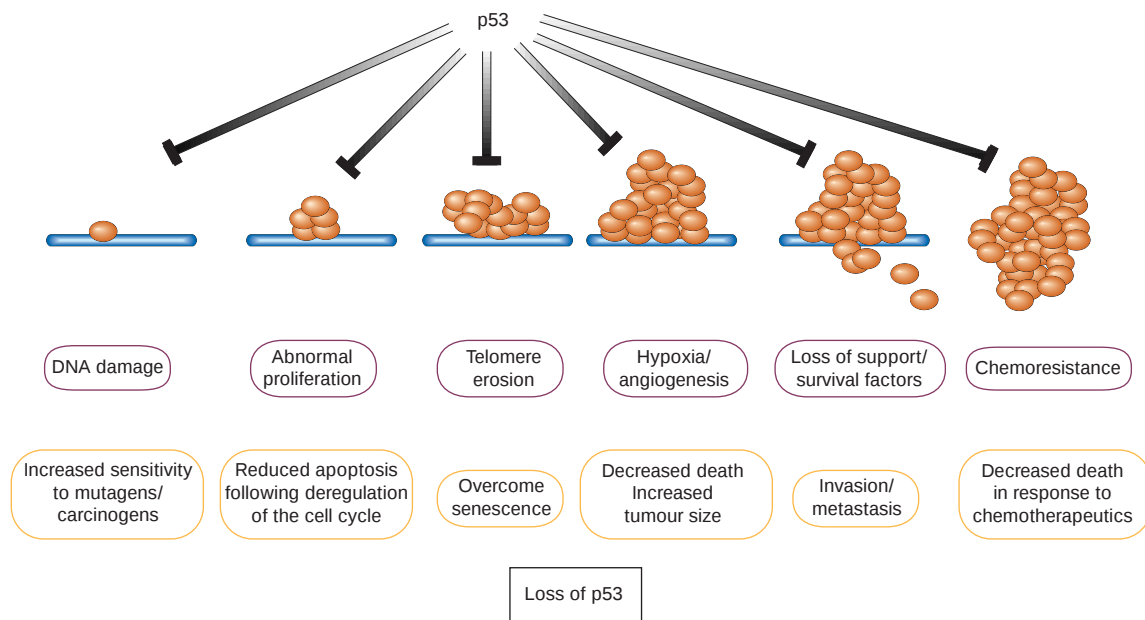


Figure 3 Many stress signals encountered during tumour progression activate p53, resulting in apoptosis or growth arrest. Loss either of the ability to activate p53 or of p53 function itself has considerable impact on the 'success' of the carcinogenic process, as it increases the chances of a tumour cell surviving progressively adverse conditions. Inability to activate p53 in response to stress signals encountered early during tumour development, such as deregulated proliferation, may be sufficient to allow the formation of preneoplastic lesions. However, lesions that suppress activation of p53 in response to such oncogene-associated stress signals do not necessarily block activation of p53 by subsequent events encountered during malignant progression, such as DNA damage. Consequently, additional alterations in pathways that activate or respond to p53, or loss of p53 by direct mutation of the gene itself, may be selected during progression to more malignant cancers.

IGF-II (ref. 32), activating mutations of Akt, a serine/threonine kinase that induces a strong survival signal^{33,34}, and loss of the suppressor of Akt function PTEN^{35–37}. The anti-apoptotic oncoproteins Bcl-2 and Bcl-x_L, which exert their principal effects through stabilization of the mitochondrion, are found to be overexpressed in several tumour types and recent analyses have indicated that loss of Apaf-1 is a relatively frequent event in malignant melanoma that presumably confers resistance to apoptosis³⁸.

A particularly potent driving force for the suppression of apoptosis in tumour cells is the coupled relationship between cell proliferation and cell death, a phenomenon exemplified by the Myc protein. In addition to its well documented growth-promoting property, Myc was found to be a powerful inducer of apoptosis, especially under conditions of stress, genotoxic damage or depleted survival factors^{39,40}. Consideration of such observations led to the proposal that the innate apoptotic potential of Myc serves as an in-built foil to its oncogenic capacity (Fig. 2 and refs 39, 41, 42). Similar antagonistic duality has since been described for essentially all known growth-promoting proteins, including E2F1 (refs 43–46), whose pro-apoptotic activity provides a counter to the proliferative effect of loss of pRB³. Even under circumstances where apoptosis is not induced by activation of oncogenes such as E2F1 (ref. 47) or Ras^{48–50}, an irreversible cell-cycle arrest is triggered in its place, which serves as an alternate mechanism to forestall continued proliferation.

Growth-deregulating oncoproteins seem to promote apoptosis through the activation of several downstream pro-apoptotic effector pathways. For example, Myc has a profound effect on the mitochondrion, triggering release of cytochrome c and activation of caspase 9. This pathway is inhibited by members of the Bcl-2/Bcl-x_L anti-apoptotic family and by survival factors, both of which have been shown to potentiate the oncogenic action of c-Myc^{51–55}. E2F1 can directly influence apoptotic signalling from death receptors⁵⁶, whereas Myc greatly enhances sensitivity to signalling through the CD95 (ref. 57), TNF⁵⁸ and TRAIL⁵⁹ death receptors. Another common pathway through which a wide variety of proliferative

signals influence the apoptotic programme is through induction of ARF, an alternate product of the *INK4a* locus, one of whose functions is to trigger upregulation of p53 through its inhibitory action on MDM-2 (ref. 60). Yet another pathway recently described for Myc seems to involve rapid downregulation of E-cadherin, which may put the affected cell into a state of *de facto* anoikis (S. Pelengaris and G.E., manuscript in preparation).

Another potent selective pressure in cancers to suppress apoptosis arises from the fact that programmed cell death is the typical response of somatic cells to many forms of stress and damage; in particular damage to cell DNA (a fact exploited by most classical cancer therapeutics). Stress-associated signals that activate apoptosis include many of those encountered by the incipient tumour cell, including hypoxia and nutrient deprivation, as well as DNA damage arising from telomere erosion, defective repair, oncogene deregulation and therapy (see review in this issue by Hoeijmakers, pages 366–374). The p53 protein is important in transducing such diverse signals into tumour-suppressive apoptotic or growth-arresting responses, which implies that there is strong selection for tumour cells to lose p53 function⁶¹. Importantly, differing p53-activating stresses tend to arise at different stages of carcinogenic progression. For example, oncogene deregulation occurs early, as it is a prerequisite for clonal expansion, whereas hypoxia is significant only after the tumour reaches macroscopic size. Consequently, p53 exerts a tumour-suppressive role at multiple stages of carcinogenic progression (Fig. 3), offering an explanation for why loss of p53 has such a profound effect on tumour development.

But the notion that p53 is a cellular superhero that functions solely to protect the organism from itself is almost certainly too simplistic. In those systems where tumour progression can be followed from pre-malignancy through to invasive cancers, p53 mutation is seldom one of the earliest events. For example, in both mouse skin carcinogenesis⁶² and human colon cancer development⁶³, mutation of p53 occurs at the point of transition from pre-malignant to invasive lesions, well after activation of some of the oncogenes that are thought

to trigger the p53 response. One probable reason for this is that alternative mutations in early-stage tumours serve to incapacitate some aspects of the p53 response. The best described of these affects ARF, whose loss severs the link between deregulation of oncoproteins such as Ras, Myc and E2F, and consequently p53 activation, permitting cells to proliferate and survive in the face of oncogene deregulation. Although mutations specifically altering the ARF protein are uncommon in human cancers, other mechanisms that hinder ARF function have been described, including methylation of the ARF promoter and amplification of genes such as *Bmi-1* (ref. 64), *Twist*⁶⁵ and *TBX2* (ref. 66), which encode repressors of ARF expression.

Inactivation of ARF through methylation of the ARF promoter occurs in both carcinomas and adenomas of the colon^{67,68}. This probably confers on colonic enterocytes the capacity to continue to proliferate despite activation of Ras, a situation that may be further exacerbated by the ability of Ras to induce expression of the p53 inactivator MDM-2 (ref. 69). But although loss of ARF serves to suppress the p53 response to oncogene activation, it leaves p53 available within the cell to respond to other ARF-independent stress. Ultimately, the evolving cancer cell will still run into a p53-induced block, at which point inactivation of p53 may be the only mechanism by which the tumour cell can endure. Of course, such a model begs the question: why is p53 not mutated in early pre-malignant lesions, as this would presumably strip the cell of any opposition to malignant progression? One possibility is that ARF possesses p53-independent tumour-suppressive activities that are independently selected against in early neoplasias. Another intriguing notion is that loss of p53 could confer some kind of immediate selective disadvantage upon the affected cell that must be overcome before the tumour can progress further. This idea is supported by surprising experimental data indicating that p53-null mice are less susceptible to development of carcinogen-induced papillomas^{70–72}. However, once neoplastic lesions do arise in such mice, albeit at greatly reduced frequency, their progression to invasive carcinoma is more or less immediate.

Not only can p53 loss have different effects at various stages of carcinogenesis, but it can also have far-reaching consequences for the evolutionary trajectory of tumour progression by transforming potent tumour-suppressive mechanisms into powerfully oncogenic ones. For example, erosion of telomeres in aberrantly proliferating cells generates a powerful DNA damage signal that triggers p53-dependent growth arrest and apoptosis, and efficiently ablates potential tumour cells that exhaust their proliferative potential. However, cells that lack functional p53 are unable to respond in this way and are forced to endure the catastrophic consequences of telomere erosion, resulting in 'rampant genome instability'^{14,73}. Similarly, oncogenic consequences of defective DNA-repair machinery are probably minimal in p53-positive cells that can respond appropriately to damaged DNA. By contrast, the combination of compromised repair (a process to which p53 also contributes) together with suppressed apoptosis is likely to constitute a heady oncogenic brew.

Restraints to the acquisition of heritable diversity

As already described, cancer development depends on the acquisition and selection of specific characteristics that set the tumour cell apart from normal somatic cells. It is thought that most cancer is precipitated by *de novo* mutations in somatic cells, a process that may be accelerated by the genomic instability inherent to most cancers⁷⁴. However, the extent to which genomic instability is a prerequisite for tumour development remains unclear, as to some degree the chromosomal chaos characteristic of almost all tumour cells may be merely be an indicator of some past acute genome-destabilizing event, such as telomere erosion. Moreover, the requirement for new mutations to drive tumour progression may be partly substituted by loss of mechanisms that limit the phenotypic expression of innate genetic variation that is inherent to all cells. Loss of HSP90, for example, has been shown to reveal extensive morphological variation that

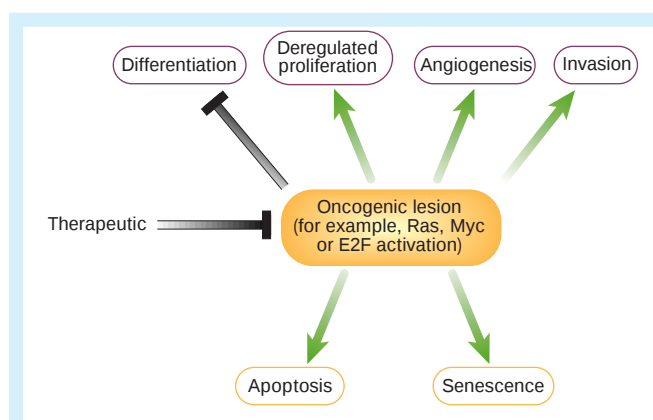


Figure 4 Growth deregulating lesions generate profound, diverse and cell-type specific pleiotropic changes in a cell and its surrounding. Some of these (proliferation, angiogenesis, suppression of terminal differentiation, local invasion) augment the neoplastic effect of the primary lesion, whereas others (sensitization to apoptosis, induction of growth arrest or senescence) are innate defences that inhibit it. Inhibiting the primary growth-deregulating lesion will influence all of these downstream sequelae. The net result is not straightforward to predict and will vary depending upon the cell type affected and the composition of lesions driving the particular neoplasm.

is usually silenced⁷⁵. The existence of protein variability that is normally buffered through protein-polishing mechanisms like HSP90 leads to the possibility that release of this innate variation may complement, and to some degree substitute for, the requirement for new somatic mutations during tumour development.

Therapeutic targeting of cell proliferation and apoptosis

Because deregulated proliferation and inhibition of apoptosis lie at the heart of all tumour development, they present two obvious targets for therapeutic intervention in all cancers. Clearly there are numerous mechanisms through which these two defects can occur, and the success of targeted therapy will depend to a large part on the molecular fingerprinting of individual tumours.

Although most existing cancer drugs are anti-mitotic, they act not by targeting the specific lesions responsible for deregulated tumour growth, but by crudely interfering with the basic machinery of DNA synthesis and cell division. Moreover, we now know that the surprising selectivity of such crude agents results largely from the increased sensitivity to apoptosis afforded to tumour cells by their oncogenic lesions^{3,39,76}. Drugs designed to specifically inhibit growth-deregulating lesions are currently being tested in clinical trials, and include inhibitors of RTKs, Ras, downstream signalling kinases such as the mitogen-activate protein kinase and Akt pathway, and CDKs⁷⁷.

At first glance, targeted inhibition of growth-deregulating lesions in cancer would seem to have limited therapeutic efficacy, as they would at best be cytostatic. However, unexpected therapeutic bonuses may emerge from such an approach because growth deregulation induces a plethora of downstream activities in affected cells and their adjacent tissues. For example, growth-deregulating lesions such as E2F and Myc are potent inhibitors of differentiation in many cell lineages. Therapeutic inhibition of the offending oncoprotein in tumours arising from cell lineages where terminal differentiation has been blocked could be sufficient to trigger a resumption of that differentiation programme, permanently expelling the tumour cell from the proliferating compartment. Such ideas receive support from several *in vivo* mouse models. For example, in skin tumours induced by deregulated Myc expression, subsequent inactivation of Myc leads not only to cessation of proliferation, but also to the expeditious resumption of normal keratinocyte differentiation which rapidly becomes irreversible⁷⁴. A similar resumption of terminal differentiation pathways is also observed after removal of the Myc signal in Myc-induced T-cell lymphomas⁷⁸.

Another direct consequence of certain oncogenic lesions is angiogenesis. Both activated Ras and deregulated Myc are potently angiogenic, suggesting that their pharmacological inhibition might foster the collapse of tumour vasculature. In a reversible Ras-dependent mouse model of melanoma, inactivation of Ras triggers the rapid involution of tumour vasculature, with concomitant regression of the tumour⁷⁹. Similarly, Myc has potent angiogenic capacity that has been observed in skin²⁴, pancreatic β cells (S. Pelengaris and G.E., unpublished data), lymphoma⁸⁰, neuroblastoma⁸¹ and in a fibroblast xenograph model⁸². Myc directly induces angiogenesis without any apparent need for an angiogenic switch, in part by induction of vascular endothelial growth factor (VEGF)²⁴ and possibly downregulation of the angiogenesis negative modulator thrombospondin-1 (ref. 83). Importantly, Myc-induced angiogenesis is of the leaky, immature and unstable kind so often associated with neoplasia. And, as seen in the Ras model system, inactivation of Myc in switchable Myc transgenic models of skin and β cells leads to rapid regression of tumour vasculature, triggering concomitant tumour involution (ref. 24, and S. Pelengaris and G.E., unpublished data).

Such studies offer encouragement for the idea of therapies based around specific targeting of the cell's proliferative machinery. However, anti-proliferative therapeutics need to be approached with caution. As outlined above, growth-deregulatory mutations trigger pleiotropic and tissue-specific effects, some of which serve to enhance the malignant state (proliferation, angiogenesis, suppression of differentiation), whereas others (sensitization to apoptosis) suppress it (Fig. 4). As these would all be inhibited by a single agent that blocks the initiating growth-deregulatory lesion, the therapeutic consequences of such an agent are likely to be highly tissue- and tumour-specific and, at present, difficult to predict.

The second obvious strategy for cancer therapy is to target the lesions that suppress apoptosis in tumour cells. The potent pro-apoptotic effects of growth-deregulating mutations mean that tumours are peculiarly dependent upon their particular suite of anti-apoptotic mutations for continued survival. Thus, although apoptosis in tumour cells is sufficiently suppressed to below a critical threshold to enable them to survive, they remain acutely sensitized to apoptosis. In most, if not all, cancer, this ability to survive results in part from inhibition of the p53 pathway, either by inactivating mutations in p53 itself, perturbation of the signalling pathways that allow activation of p53 in response to stress, or defects in the downstream mediators of p53-induced apoptosis. Reintroduction of p53 function is sufficient to induce apoptosis in many tumour cells, and several mechanisms to reactivate p53 are being considered as therapeutic strategies. These include introduction of wild-type p53 into tumours expressing a mutant protein, or inhibition of negative regulators of p53, such as MDM-2, in those tumours that retain wild-type p53 (ref. 61).

Interference with survival signalling is another appealing approach to the induction of apoptosis in tumour cells, either by direct inhibition of components of the signalling cascades, such as STI571 inhibition of Bcr-Abl in chronic myelogenous leukaemia⁸⁴, or by inhibition of angiogenesis by drugs that target the VEGF receptors Flt-1 and KDR⁸⁵. Reintroduction of inhibitors of VEGF expression, such as VHL, also represent interesting targets in this context⁸⁶. Direct participants of apoptotic pathways, such as the Bcl-2 proteins that are important in both cancer development and the acquisition of resistance to conventional cancer therapies, provide further targets for the development of drugs that may be indifferent to the p53 status of the tumour cell⁸⁷.

Regardless of efficiency in cell killing, the success of repairing the apoptotic response in tumour cells depends on the extent to which such therapies confine death to the cancer cells, and allow survival of normal tissue. Many conventional chemotherapies induce significant toxicity, particularly in tissues that normally maintain a proliferative compartment, such as gut epithelium and the haematopoietic system. This DNA damage-induced toxicity is mediated in part through p53, leading to the suggestion that inhibition of

p53 in these normal tissues may protect against drug-induced toxicity, thereby improving the tolerance of conventional cancer therapies⁸⁸. However, implicit in the development of drugs that target specific lesions responsible for tumour cell growth is the prediction that these approaches will show significantly more specificity for tumour cell killing than conventional therapies.

Although activation of apoptotic pathways can lead to the death of untransformed cells, a process that is essential in normal development, a fundamental difference exists between tumour cells and their normal counterparts, as normal cells neither have to sustain the pro-apoptotic onslaught that is inherent in deregulated proliferation, nor survive away from their usual environment in the absence of requisite survival signals. Repair or replacement of a single apoptotic signal, be it reactivation of p53 or removal of a survival signal, could well prove too much for a tumour cell already burdened with a heavy apoptotic load. By contrast, the same perturbation may scarcely ruffle the equilibrium of a normal cell, safely buffered in its appropriate soma and enjoying the full gamut of trophic support that ensures normal cell survival. An interesting variation on this theme is illustrated by the activity of antagonists of Cdk2. These inhibitors, which would ostensibly function to prevent cell-cycle progression, prevent normal phosphorylation and inactivation of E2F1 at the completion of DNA synthesis. The outcome is tumour-specific apoptosis, presumably stemming from an inability of tumour cells to tolerate yet further deregulation of E2F activity, beyond that already sustained through perturbation of the pRB pathway⁸⁹. Whether this difference between normal and tumour cells actually exists in a meaningful way, and whether we can fully exploit it in the development of new drugs to treat cancers, are questions and challenges that now face us.

Clearly, all forms of tumour therapy carry with them the danger of selection for resistance, a problem that may be exacerbated by the genomic plasticity inherent in most, if not all, cancers. The most effective solution to this problem is almost certainly to simultaneously attack multiple lesions specific to individual tumours, in a much more sophisticated version of standard combined chemotherapies used at present. Evolution of cancer therapy is likely to remain a combination of design and error, but the development of mechanisms to target the mission-critical events that are common to all cancers provides a glimpse of therapeutic potential hitherto unimaginable. □

- Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
- Oller, A. R., Rastogi, P., Morgenthaler, S. & Thilly, W. G. A statistical model to estimate variance in long term-low dose mutation assays: testing of the model in a human lymphoblastoid mutation assay. *Mutat. Res.* **216**, 149–161 (1989).
- Evan, G. & Littlewood, T. A matter of life and cell death. *Science* **281**, 1317–1322 (1998).
- Pardee, A. B. G. events and regulation of cell proliferation. *Science* **246**, 603–608 (1989).
- Roovers, K. & Assoian, R. K. Integrating the MAP kinase signal into the G1 phase cell cycle machinery. *BioEssays* **22**, 818–826 (2000).
- Massague, J., Blain, S. W. & Lo, R. S. TGF β signaling in growth control, cancer, and heritable disorders. *Cell* **103**, 295–309 (2000).
- Sangfelt, O., Erickson, S. & Grandér, D. Mechanisms of interferon-induced cell cycle arrest. *Front. Biosci.* **5**, D479–D487 (2000).
- Zhu, L. & Skoultschi, A. I. Coordinating cell proliferation and differentiation. *Curr. Opin. Genet. Dev.* **11**, 91–97 (2001).
- Bach, S. P., Renahan, A. G. & Potten, C. S. Stem cells: the intestinal stem cell as a paradigm. *Carcinogenesis* **21**, 469–476 (2000).
- Booth, C. & Potten, C. S. Gut instincts: thoughts on intestinal epithelial stem cells. *J. Clin. Invest.* **105**, 1493–1499 (2000).
- Fuchs, E. & Segre, J. A. Stem cells: a new lease on life. *Cell* **100**, 143–155 (2000).
- Raff, M. C. Social controls on cell survival and cell death. *Nature* **356**, 397–400 (1992).
- Raff, M. et al. Programmed cell death and the control of cell survival: lessons from the nervous system. *Science* **262**, 695–700 (1993).
- DePinho, R. A. The age of cancer. *Nature* **408**, 248–254 (2000).
- Hunter, T. Signaling—2000 and beyond. *Cell* **100**, 113–127 (2000).
- Harbour, J. W. & Dean, D. C. The Rb/E2F pathway: expanding roles and emerging paradigms. *Genes Dev.* **14**, 2393–2409 (2000).
- Sherr, C. J. Cancer cell cycles. *Science* **274**, 1672–1677 (1996).
- Baudino, T. A. & Cleveland, J. L. The Max network gone mad. *Mol. Cell. Biol.* **21**, 691–702 (2001).
- Stocker, H. & Hafen, E. Genetic control of cell size. *Curr. Opin. Genet. Dev.* **10**, 529–535 (2000).
- Johnston, L. A., Prober, D. A., Edgar, B. A., Eisenman, R. N. & Gallant, P. Drosophila myc regulates cellular growth during development. *Cell* **98**, 779–790 (1999).
- Iritani, B. M. & Eisenman, R. N. c-Myc enhances protein synthesis and cell size during B lymphocyte development. *Proc. Natl Acad. Sci. USA* **96**, 13180–13185 (1999).

22. Elend, M. & Eilers, M. Cell growth: downstream of Myc—to grow or to cycle? *Curr. Biol.* **9**, R936–R938 (1999).
23. Dang, C. V. *et al.* Function of the c-Myc oncogenic transcription factor. *Exp. Cell Res.* **253**, 63–77 (1999).
24. Pelengaris, S., Littlewood, T., Khan, M., Elia, G. & Evan, G. Reversible activation of c-Myc in skin: induction of a complex neoplastic phenotype by a single oncogenic lesion. *Mol. Cell* **3**, 565–577 (1999).
25. Pelengaris, S., Rudolph, B. & Littlewood, T. Action of Myc in vivo—proliferation and apoptosis. *Curr. Opin. Genet. Dev.* **10**, 100–105 (2000).
26. Gu, W. *et al.* Interaction of myogenic factors and the retinoblastoma protein mediates muscle cell commitment and differentiation. *Cell* **72**, 309–324 (1993).
27. Lasorella, A., Nosedà, M., Beyna, M. & Iavarone, A. Id2 is a retinoblastoma protein target and mediates signalling by Myc oncoproteins. *Nature* **407**, 592–598 (2000).
28. Walczak, H. & Krammer, P. H. The CD95 (APO-1/Fas) and the TRAIL (APO-2L) apoptosis systems. *Exp. Cell Res.* **256**, 58–66 (2000).
29. Vander Heiden, M. G. *et al.* Outer mitochondrial membrane permeability can regulate coupled respiration and cell survival. *Proc. Natl Acad. Sci. USA* **97**, 4666–4671 (2000).
30. Hengartner, M. O. The biochemistry of apoptosis. *Nature* **407**, 770–776 (2000).
31. Puthalakath, H., Huang, D. C., O'Reilly, L. A., King, S. M. & Strasser, A. The proapoptotic activity of the Bcl-2 family member Bim is regulated by interaction with the dynein motor complex. *Mol. Cell* **3**, 287–296 (1999).
32. Yu, H. & Rohan, T. Role of the insulin-like growth factor family in cancer development and progression. *J. Natl Cancer Inst.* **92**, 1472–1489 (2000).
33. Datta, S. R., Brunet, A. & Greenberg, M. E. Cellular survival: a play in three Akts. *Genes Dev.* **13**, 2905–2927 (1999).
34. Stambolic, V., Mak, T. W. & Woodgett, J. R. Modulation of cellular apoptotic potential: contributions to oncogenesis. *Oncogene* **18**, 6094–6103 (1999).
35. Machama, T. & Dixon, J. E. PTEN: a tumour suppressor that functions as a phospholipid phosphatase. *Trends Cell Biol.* **9**, 125–128 (1999).
36. Bonneau, D. & Longy, M. Mutations of the human PTEN gene. *Hum. Mutat.* **16**, 109–122 (2000).
37. Kandel, E. S. & Hay, N. The regulation and activities of the multifunctional serine/threonine kinase Akt/PKB. *Exp. Cell Res.* **253**, 210–229 (1999).
38. Soengas, M. S. *et al.* Inactivation of the apoptosis effector Apaf-1 in malignant melanoma. *Nature* **409**, 207–211 (2001).
39. Evan, G. *et al.* Induction of apoptosis in fibroblasts by c-myc protein. *Cell* **63**, 119–125 (1992).
40. Askew, D., Ashmun, R., Simmons, B. & Cleveland, J. Constitutive c-myc expression in IL-3-dependent myeloid cell line suppresses cycle arrest and accelerates apoptosis. *Oncogene* **6**, 1915–1922 (1991).
41. Harrington, E. A., Fanidi, A. & Evan, G. I. Oncogenes and cell death. *Curr. Opin. Genet. Dev.* **4**, 120–129 (1994).
42. Evan, G. & Littlewood, T. The role of c-myc in cell growth. *Curr. Opin. Genet. Dev.* **3**, 44–49 (1993).
43. Almasan, A. *et al.* Deficiency of retinoblastoma protein leads to inappropriate S-phase entry, activation of E2F-responsive genes, and apoptosis. *Proc. Natl Acad. Sci. USA* **92**, 5436–5440 (1995).
44. Qin, X. Q., Livingston, D. M., Kaelin, W. G. Jr & Adams, P. D. Deregulated transcription factor E2F-1 expression leads to S-phase entry and p53-mediated apoptosis. *Proc. Natl Acad. Sci. USA* **91**, 10918–10922 (1994).
45. Shan, B. & Lee, W. H. Deregulated expression of E2F-1 induces S-phase entry and leads to apoptosis. *Mol. Cell Biol.* **14**, 8166–8173 (1994).
46. Wu, X. & Levine, A. J. p53 and E2F-1 cooperate to mediate apoptosis. *Proc. Natl Acad. Sci. USA* **91**, 3602–3606 (1994).
47. Dimri, G. P., Itahana, K., Acosta, M. & Campisi, J. Regulation of a senescence checkpoint response by the E2F1 transcription factor and p14(ARF) tumor suppressor. *Mol. Cell Biol.* **20**, 273–285 (2000).
48. Hirakawa, T. & Ruley, H. E. Rescue of cells from ras oncogene-induced growth arrest by a second, complementing, oncogene. *Proc. Natl Acad. Sci. USA* **85**, 1519–1523 (1988).
49. Ferbeyre, G. *et al.* PML is induced by oncogenic ras and promotes premature senescence. *Genes Dev.* **14**, 2015–2027 (2000).
50. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. *Cell* **88**, 593–602 (1997).
51. Fanidi, A., Harrington, E. & Evan, G. Cooperative interaction between c-myc and bcl-2 proto-oncogenes. *Nature* **359**, 554–556 (1992).
52. Bissonnette, R., Echeverri, F., Mahboubi, A. & Green, D. Apoptotic cell death induced by c-myc is inhibited by bcl-2. *Nature* **359**, 552–554 (1992).
53. Wagner, A. J., Small, M. B. & Hay, N. Myc-mediated apoptosis is blocked by ectopic expression of bcl-2. *Mol. Cell Biol.* **13**, 2432–2440 (1993).
54. Harrington, E. A., Bennett, M. R., Fanidi, A. & Evan, G. I. c-Myc-induced apoptosis in fibroblasts is inhibited by specific cytokines. *EMBO J.* **13**, 3286–3295 (1994).
55. Strasser, A., Harris, A. W., Bath, M. L. & Cory, S. Novel primitive lymphoid tumours induced in transgenic mice by cooperation between myc and bcl-2. *Nature* **348**, 331–333 (1990).
56. Phillips, A. C., Ernst, M. K., Bates, S., Rice, N. R. & Vousden, K. H. E2F-1 potentiates cell death by blocking anti-apoptotic signaling pathways. *Mol. Cell* **4**, 771–781 (1999).
57. Hueber, A.-O. *et al.* Requirement for the CD95 receptor-ligand pathway in c-Myc induced apoptosis. *Science* **278**, 1305–1309 (1997).
58. Klefstrom, J. *et al.* c-Myc induces cellular susceptibility to the cytotoxic action of TNF- α . *EMBO J.* **13**, 5442–5450 (1994).
59. Lutz, W., Fulda, S., Jeremias, I., Debatin, K. M. & Schwab, M. MycN and IFN γ cooperate in apoptosis of human neuroblastoma cells. *Oncogene* **17**, 339–346 (1998).
60. Sherr, C. J. & Weber, J. D. The ARF/p53 pathway. *Curr. Opin. Genet. Dev.* **10**, 94–99 (2000).
61. Woods, D. B. & Vousden, K. H. Regulation of p53 function. *Exp. Cell Res.* **264**, 56–66 (2001).
62. Frame, S. *et al.* Epithelial carcinogenesis in the mouse: correlating the genetics and the biology. *Phil. Trans. R. Soc. Lond. B* **353**, 839–845 (1998).
63. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
64. Jacobs, J. J., Kieboom, K., Marino, S., DePinho, R. A. & van Lohuizen, M. The oncogene and Polycomb-group gene *bmi-1* regulates cell proliferation and senescence through the *ink4a* locus. *Nature* **397**, 164–168 (1999).
65. Maestros, R. *et al.* *twist* is a potential oncogene that inhibits apoptosis. *Genes Dev.* **13**, 2207–2217 (1999).
66. Jacobs, J. J. *et al.* Senescence bypass screen identifies *TBX2*, which represses *Cdkn2a* (*p19^{ARF}*) and is amplified in a subset of human breast cancers. *Nature Genet.* **26**, 291–299 (2000).
67. Esteller, M. *et al.* Hypermethylation-associated inactivation of p14(ARF) is independent of p16(INK4a) methylation and p53 mutational status. *Cancer Res.* **60**, 129–133 (2000).
68. Robertson, K. D. & Jones, P. A. The human ARF cell cycle regulatory gene promoter is a CpG island which can be silenced by DNA methylation and down-regulated by wild-type p53. *Mol. Cell Biol.* **18**, 6457–6473 (1998).
69. Ries, S. *et al.* Opposing effects of Ras on p53: transcriptional activation of mdm2 and induction of p19ARF. *Cell* **103**, 321–330 (2000).
70. Kemp, C. J., Donehower, L. A., Bradley, A. & Balmain, A. Reduction of p53 gene dosage does not increase initiation or promotion but enhances malignant progression of chemically induced skin tumors. *Cell* **74**, 813–822 (1993).
71. Greenhalgh, D. A., Wang, X. J., Donehower, L. A. & Roop, D. R. Paradoxical tumor inhibitory effect of p53 loss in transgenic mice expressing epidermal-targeted v-rasHa, v-fos, or human transforming growth factor alpha. *Cancer Res.* **56**, 4413–4423 (1996).
72. Wang, X. J., Greenhalgh, D. A., Donehower, L. A. & Roop, D. R. Cooperation between Ha-ras and fos or transforming growth factor alpha overcomes a paradoxical tumor-inhibitory effect of p53 loss in transgenic mouse epidermis. *Mol. Carcinogenesis* **29**, 67–75 (2000).
73. Counter, C. M. *et al.* Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *EMBO J.* **11**, 1921–1929 (1992).
74. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).
75. Rutherford, S. L. & Lindquist, S. Hsp90 as a capacitor for morphological evolution. *Nature* **396**, 336–342 (1998).
76. Schmitt, C. A. & Lowe, S. W. Apoptosis and therapy. *J. Pathol.* **187**, 127–137 (1999).
77. Gibbs, J. B. Mechanism-based target identification and drug discovery in cancer research. *Science* **287**, 1969–1973 (2000).
78. Felsher, D. W. & Bishop, J. M. Reversible tumorigenesis by myc in hematopoietic lineages. *Mol. Cell* **4**, 199–207 (1999).
79. Chin, L. *et al.* Essential role for oncogenic Ras in tumour maintenance. *Nature* **400**, 468–472 (1999).
80. Brandvold, K. A., Neiman, P. & Ruddell, A. Angiogenesis is an early event in the generation of myc-induced lymphomas. *Oncogene* **19**, 2780–2785 (2000).
81. Breit, S. *et al.* The N-myc oncogene in human neuroblastoma cells: down-regulation of an angiogenesis inhibitor identified as activin A. *Cancer Res.* **60**, 4596–4601 (2000).
82. Ngo, C. V. *et al.* An in vivo function for the transforming Myc protein: elicitation of the angiogenic phenotype. *Cell Growth Differ.* **11**, 201–210 (2000).
83. Janz, A., Sevignani, C., Kenyon, K., Ngo, C. V. & Thomas-Tikhonenko, A. Activation of the myc oncoprotein leads to increased turnover of thrombospondin-1 mRNA. *Nucleic Acids Res.* **28**, 2268–2275 (2000).
84. O'Dwyer, M. E. & Druker, B. J. Status of bcr-*abl* tyrosine kinase inhibitors in chronic myelogenous leukemia. *Curr. Opin. Oncol.* **12**, 594–597 (2000).
85. Morin, M. J. From oncogene to drug: development of small molecule tyrosine kinase inhibitors as anti-tumor and anti-angiogenic agents. *Oncogene* **19**, 6574–6583 (2000).
86. Krek, W. VHL takes HIF's breath away. *Nature Cell Biol.* **2**, E1–E3 (2000).
87. Huang, Y. Q., Li, J. J. & Karparkin, S. Thrombin inhibits tumor cell growth in association with up-regulation of p21(waf/cip1) and caspases via a p53-independent, STAT-1-dependent pathway. *J. Biol. Chem.* **275**, 6462–6468 (2000).
88. Komarov, P. G. *et al.* A chemical inhibitor of p53 that protects mice from the side effects of cancer therapy. *Science* **285**, 1733–1737 (1999).
89. Chen, Y. N. *et al.* Selective killing of transformed cells by cyclin/cyclin-dependent kinase 2 antagonists. *Proc. Natl Acad. Sci. USA* **96**, 4325–4329 (1999).

The Hedgehog and Wnt signalling pathways in cancer

Jussi Taipale & Philip A. Beachy

Department of Molecular Biology and Genetics, Howard Hughes Medical Institute, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA (e-mail: pbeachy@jhmi.edu)

The Wnt and Hedgehog (Hh) signalling pathways have long been known to direct growth and patterning during embryonic development. Recent evidence also implicates these pathways in the postembryonic regulation of stem-cell number in epithelia such as those of the skin and intestine, which undergo constant renewal. A pathological role for the Wnt and Hh pathways has emerged from studies showing a high frequency of specific human cancers associated with mutations that constitutively activate the transcriptional response of these pathways. This article focuses on Hh and Wnt signal transduction and reviews evidence suggesting that tumorigenesis associated with pathway activation may result from mis-specification of cells towards stem-cell or stem cell-like fates.

Single-celled organisms grow and divide constrained only by the availability of nutrients in the environment. In contrast, the growth of animal cells is directed by mechanisms that have evolved to establish and maintain the optimal size and function of interdependent organs. Cancer cells subvert these evolutionary adaptations to multicellularity and revert to a largely nutrient-limited style of growth. From this perspective, an understanding of the mechanisms of normal growth control should help explain the deranged cell growth associated with cancer. But it is only in the past ten years that the genetics of development and cancer have converged in the identification of extracellular signalling pathways that are aberrantly regulated in cancer and are also central to embryonic patterning (Table 1). This article focuses on general aspects common to pathways headed by the Hh and Wnt families of secreted signalling proteins, and on the mechanisms by which aberrant regulation of these pathways can contribute to the tumour phenotype (for more comprehensive reviews of the individual pathways, see refs 1–7).

The tumours associated with the Wnt and Hh signalling pathways arise typically from tissues in which the pathways normally operate. In tumours, however, oncogenic mutations lock these pathways into ligand-independent states of constitutive activity. Activation of these pathways does not function merely as a mitogenic stimulus, as there is no simple correlation in embryos between pathway activity and cell proliferation. In tumours, therefore, activation of these pathways might be thought of as causing mis-specification of cells within the context of the target organ towards a fate permitting proliferation. But the tumours associated with these pathways are adult as well as paediatric. How can mis-specification of cells in a fully formed adult tissue contribute to tumorigenesis? In this context it is important that epithelia, which are the source of 90% of adult tumours, require constant renewal to maintain their integrity. As this renewal must be governed by growth-regulatory mechanisms, pattern specification continues beyond embryogenesis (Fig. 1). The linings of the intestine and skin, for example, are fast-renewing tissues (necessitated by constant environmental exposure), yet these epithelia maintain a precisely patterned organization throughout

life. Tumours arising from these epithelia are associated with aberrant regulation of the Wnt and Hh pathways, which are important in regulating their normal growth and patterning.

Hedgehog signal transduction

The *hh* gene was identified as a secreted signalling protein required for specification of positional identity in the *Drosophila* embryonic segment^{4,5,7–9}. The Hh protein in *Drosophila* also controls patterning of imaginal disc-derived adult structures such as the eye and appendages, and of the abdominal cuticle. The three mammalian *hh* genes, *Sonic*, *Indian* and *Desert hedgehog* (*Shh*, *Ihh* and *Dhh* respectively)^{4–7}, are important in the patterning of many tissues and structures; loss or reduction of Hh signalling is associated with numerous developmental deficits and malformations, one of the most striking of which is the cyclopia associated with loss of *Shh* signalling⁶.

Hh proteins enter the secretory pathway and undergo an intramolecular cleavage and lipid modification reaction catalysed by the carboxy-terminal portion of the precursor. This results in an amino-terminal peptide of relative molecular mass 19,000, esterified at its C terminus to a cholesterol molecule (HhNp)¹⁰. This peptide is responsible for all known signalling activities of the Hh protein. The mammalian *ShhNp* protein undergoes further palmitoylation at its N terminus¹¹; efficient addition of palmitate is dependent upon prior cholesterol addition, and may enhance signalling activity of the protein in some settings. Despite the presence of two lipophilic moieties, *ShhNp* seems to directly influence distant cells in the developing vertebrate limb and neural tube. In *Drosophila*, cholesterol modification seems to target the Hh signal to a delivery system that includes Dispatched, a protein related to Patched and required for release of Hh from the secreting cell, and Tout velu, an enzyme involved in heparan sulphate biosynthesis that is required for efficient transport of, and response to, the Hh protein signal^{12,13}.

In contrast to most signalling pathways, intracellular Hh signal transduction proceeds largely by sequential repressive interactions (Fig. 2). Response to the Hh signal is controlled by two transmembrane proteins, the tumour-suppressor Patched (Ptc) and the proto-oncogene Smoothened (Smo)^{4,5,9}. Ptc is a twelve-span transmem-

brane protein structurally similar to the putative proton-driven lipid translocator mutated in Niemann-Pick C1 disease^{5,14}. Smo is a member of the seven transmembrane-receptor family, most closely related to the Frizzled family of Wnt receptors (see below). In the resting state, Smo activity is suppressed by Ptc; Hh stimulation releases this inhibition, leading to Smo activation of a transcriptional response^{4,9}. Biochemical studies have provided evidence for a physical interaction between Hh and Ptc and between Ptc and Smo proteins, leading to the proposal of a heteromeric receptor model¹⁵ in which Hh binding to Ptc within the Ptc/Smo complex releases Smo activity without dissociation of Ptc and Smo. Although recent studies have substantiated a physiologically meaningful interaction between Hh and Ptc, they have shown distinct localizations for Ptc and Smo *in vivo*, suggesting that other models of receptor function should be considered^{9,16}.

It is not well understood how activation of Smo^{16,17} is coupled to the cytoplasmic proteins involved in Hh signalling, which include the serine/threonine protein kinase Fused (Fu), Suppressor of Fused (Su(fu)), the kinesin-like protein Costal-2 (Cos2), and the transcription factor Cubitus interruptus (Ci; Gli in mammals)⁴. These proteins form a large cytoplasmic complex that is anchored to microtubules, apparently through Cos2 (ref. 5). In the absence of Hh, Ci (Ci155) is phosphorylated by protein kinase A and subsequently processed to generate an N-terminal transcriptional repressor (Ci75)¹⁸. Upon Hh stimulation, the large cytoplasmic complex dissociates from the microtubules¹⁹, and the full-length Ci transcriptional activator is translocated to the nucleus²⁰, leading to transcriptional activation of Hh target genes. One of these target genes is *Ptc*, resulting in feedback inhibition²¹. Not all mammalian homologues to the *Drosophila* cytoplasmic proteins have been identified, and some homologues are present in multiple isoforms. For example, the transcriptional regulatory functions of Ci seem to be executed by Gli3, primarily a repressor, and Gli1 and Gli2, primarily activators²².

Wnt signal transduction

The Wnt signalling pathway is involved in a wide range of embryonic patterning events (reviewed in refs 1–3). One of the most striking effects of Wnt proteins is their ability to induce formation of a new embryonic axis in metazoans ranging from *Hydra* to *Xenopus*²³. In *Drosophila*, generation of an active Wingless signal requires the activity of Porcupine, a polytopic membrane protein related to acyltransferases¹. Wnt signal response seems to be facilitated by an accessory receptor, Dally, a glypican-type heparan sulphate proteoglycan¹², which is linked to the plasma membrane by a glycosylphosphatidylinositol moiety.

The signal-transducing components of the Wnt receptor are members of the low-density lipoprotein receptor-related protein (LRP)²⁴ and Frizzled²⁵ protein families. In the absence of pathway stimulation, β -catenin protein is destabilized by a cytoplasmic complex containing the proteins Axin, adenomatous polyposis coli (APC), and glycogen synthase kinase-3 β (GSK-3 β)³. The action of this complex is antagonized by Dishevelled, a cytoplasmic protein that is activated by an unknown mechanism upon binding of Wnt to its receptor. Wnt signalling thus stabilizes β -catenin^{1,3}, which acts as a transcriptional co-activator by associating with the Tcf/LEF family of transcription factors^{3,26}. In the absence of pathway stimulation, reduced β -catenin levels permit repression of Wnt target genes by association of transcriptional co-repressors with Tcf/LEF^{2,3}.

Accumulation of β -catenin is therefore critical for activation of the Wnt transcriptional response. Full activation of the Hh transcriptional response also requires the degradation of the Ci/Gli repressor. Critical roles for the relatively slow processes of protein synthesis or degradation in the activation of these pathways represents an opportunity to integrate signal strength over time. The increased precision inherent in this type of signal response may be particularly useful in

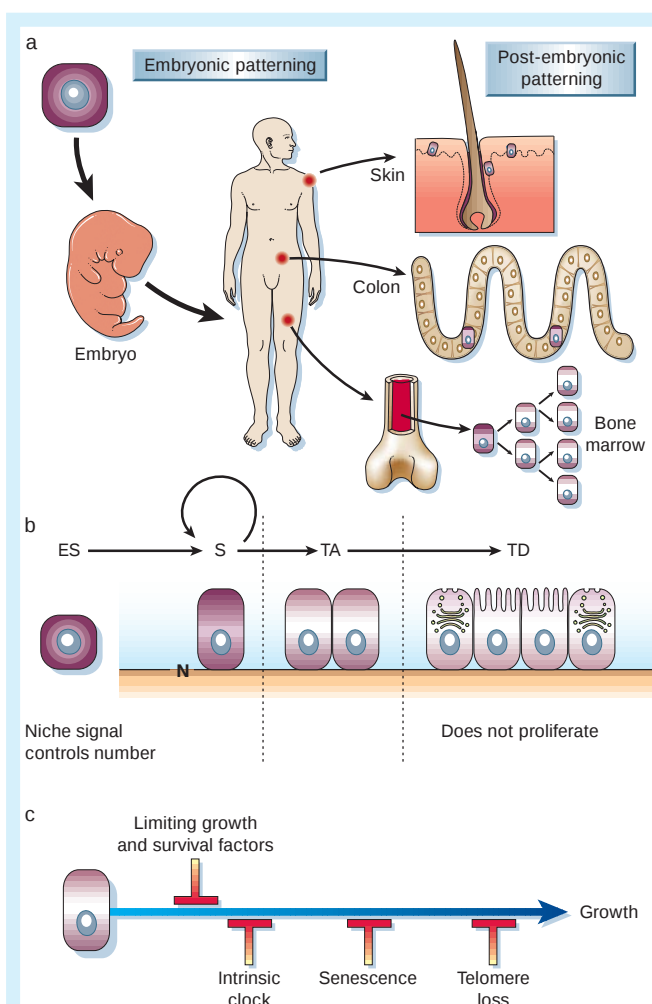


Figure 1 Development and growth control. **a**, Embryonic and postembryonic patterning. Embryonic patterning defines the identity and position of organs (left). Most adult tumours originate from tissues such as skin, intestine or blood (right), where somatic stem cells (dark purple) persist and allow patterning and growth to continue throughout life. Physiological growth control is poorly understood⁶⁵. **b**, Organizational hierarchy of growth and differentiation. Totipotent stem cells (ES) give rise to organ- or tissue-specific stem cells (S), whose proliferation in tissues is controlled by an interaction between the stem cell and its microenvironment (niche, N), which permits only a limited number of cells to retain stem-cell identity^{44–46}. Stem-cell progeny that do not receive the niche signal differentiate to transit amplifying cells (TA), which divide a limited number of times to give rise to terminally differentiated cells (TD). **c**, Limits to growth of transit amplifying cells. Growth of transit amplifying cells prior to terminal differentiation is limited by extrinsic survival and growth signals^{53,55,65}. Loss of critical extrinsic factors leads to cell death or immediate terminal differentiation. These extrinsic signals can lead to an increase in the number of times the transit amplifying cells proliferate, but they apparently cannot support proliferation indefinitely. The time the cells can divide before terminally differentiating seems to be limited by a clock mechanism involving cyclin-dependent kinase inhibitors (for example, p27 in mice, Dacapo in *Drosophila*)^{55,65}. Two other mechanisms relevant to cancer but probably not to normal growth control are accumulation of the cyclin-dependent kinase inhibitor p16^{INK4a} ('senescence') and telomere loss ('crisis'⁶⁶).

the context of embryonic development, where the degree of pathway activation within cells at varying distances from a localized signal determines the pattern of structures ultimately formed.

The genetics of aberrant Wnt and Hh pathway activation

Early evidence for involvement of the Wnt pathway in cancer came from isolation of *Wnt-1* as *Int-1*, a gene activated by nearby integration of the mouse mammary tumour virus in a mammary tumour

Table 1 Wnt and Hedgehog pathways in cancer

Pathway	Tumour type	Occurrence of mutations in sporadic cases	Familial syndrome, tumour incidence
Hedgehog	Basal cell carcinoma	~50%	BCNS, ~100%
	Medulloblastoma	~25%	BCNS, 1–3%
	Fibrosarcoma	ND	BCNS, low
	Rhabdomyosarcoma	ND	BCNS, very low
Wnt	Colorectal cancer	85%	FAP, very high in untreated cases
	Desmoid tumour	74%	FAP, 10%
	Hepatoblastoma	67%	FAP, <1%

The list presented is not comprehensive and underestimates the prevalence of mutations, as neither all components of a given pathway nor transcriptional targets indicative of pathway activation have generally been examined. Included are cases where clear genetic evidence links increased cancer risk in humans or mice to a germline loss of function of a single copy of a tumour suppressor (*PTCH* in Hh, *APC* in Wnt). ND, no data; BCNS, basal cell nevus syndrome; FAP, familial adenomatous polyposis. (Source: OMIM (<http://www.ncbi.nlm.nih.gov/omim/>) and refs 2, 32, 33, 35, 47, 68–70.)

model²⁷. Later, it was found that *APC* is the tumour suppressor in familial adenomatous polyposis, a hereditary syndrome associated with a substantial increase in risk of colorectal and other cancers (Table 1). Mutational loss of *APC* function activates the Wnt transcriptional response by stabilizing β -catenin. Most sporadic colorectal tumours also involve constitutive activation of the Tcf-mediated Wnt transcriptional response, due either to loss of *APC* or to stabilizing oncogenic mutations in β -catenin^{3,28}. Transcriptional targets include genes that regulate positional specification, such as the caudal-related homeodomain protein *Cdx1* (ref. 29), and genes more directly involved in growth, such as *c-Myc*³⁰. Although it has not been formally demonstrated that the transcriptional activator function of β -catenin is its only oncogenic function, genetic manipulations that perturb β -catenin nuclear localization lead to a loss of proliferative capacity of colon cancer cells³¹.

The involvement of the Hh pathway in human cancer was appreciated upon identification of *PTCH* as the tumour suppressor in Gorlin's syndrome, which is associated with an increased risk of basal cell carcinoma (BCC), medulloblastoma and other tumours^{32,33}. Loss of *PTCH* function is also seen in sporadic tumours of the same types (Table 1). In addition, missense mutations in *SMO* were identified in sporadic tumours and shown to constitutively activate the pathway in the absence of Hh^{34,35}. A remarkable feature of familial Hh and Wnt pathway mutations is the extremely high incidence of certain tumour types, with a very high incidence of colorectal cancer in *APC* heterozygotes and a virtual certainty of developing multiple BCC in *PTCH* heterozygotes (Table 1). Although mis-expression of the Gli1 transcriptional effector causes BCC-like tumours in mice³⁶, and so establishes the importance of the transcriptional response in tumorigenesis, blocking Gli-based transcription has not yet been shown to arrest tumour growth.

Other signalling pathways important in embryonic pattern formation include the Notch pathway and the tyrosine kinase receptor/Ras pathways (reviewed in ref. 37; see article by Blume-Jensen and Hunter in this issue, pages 355–365), and those headed by members of the transforming growth factor (TGF)- β superfamily. Mutations constitutively activating Ras signalling are quite common in human tumours, and activation of Notch pathway is linked to a subset of acute T-cell lymphoblastic leukaemias. However, no mutations that activate TGF- β pathways have been reported in human tumours. In contrast, multiple forms of human cancer are associated with pathway-inactivating mutations in the TGF- β type I and II receptors and in the intracellular signal transducers Smad2 and Smad4. These loss-of-function mutations lead to a loss of growth-inhibitory responses to TGF- β 1–3, and are thought to be important in tumour progression^{38,39}.

A stem-cell connection for Hh and Wnt pathways in cancer?

What might be the connection between tumours and the Hh and Wnt pathways and how does pathway activation lead in some cases to such highly efficient tumorigenesis? Recent genetic evidence suggests: (1) that somatic stem cells are the locus of tumour initiation;

(2) that the Wnt and Hh pathways function in the normal regulation of stem-cell number in at least some tissues; and (3) that expansion of the somatic stem-cell population may be the first step in formation of at least some types of cancer.

Numerous arguments suggest a stem-cell origin for human cancer. First, it is worth noting that stem cells possess many of the features that constitute the tumour phenotype, including self-renewal and essentially unlimited replicative potential⁴⁰. Second, the mutations that initiate tumour formation seem to accumulate in cells that persist throughout life, as suggested by the exponential increase of cancer incidence with age. This is thought to reflect a requirement for between four and seven mutations in a single cell to effect malignant transformation⁴⁰. Similarly, cancer formation from cells that persist throughout life is suggested by an increased incidence in adults of skin tumours such as melanoma after higher childhood exposure to a mutagenic agent such as ultraviolet solar radiation⁴¹. Normal somatic stem cells are strong candidates for such persistent cells. An alternative would be that mutation within a more differentiated cell might break the normal growth-regulatory mechanisms that limit proliferative capacity (Fig. 1c) and result in a persistent clone of proliferating cells. This seems unlikely, at least for sporadic tumours caused by loss of tumour suppressors such as *PTCH* or *APC*, because non-stem cells are generally destined for terminal differentiation within a time window too short for acquisition of sequential mutations that must affect two copies of a wild-type tumour suppressor (Fig. 3).

Stem cell-derived clones of cells lacking function of the p53 protein have been found in phenotypically normal sun-exposed skin⁴². The p53 protein acts in maintenance of genome integrity, and its loss in stem cells in the absence of tissue dysplasia indicates that the stem-cell compartment could indeed be the target of genetic events in the earliest stages of tumorigenesis (Fig. 3). In fact, many tumours seem to derive from multipotential progenitor cells, as indicated by the presence of multiple clonally related cell types within the tumour. This is particularly apparent in leukaemias, perhaps because the ratio of differentiated cells to stem cells in the haematopoietic system is high. Ample evidence indicates a haematopoietic stem-cell origin for chronic myeloid leukaemia. And a stem cell-like origin for acute myeloid leukaemia (AML) is suggested by tumour transplantation studies demonstrating that only 0.2–100 in 10⁶ white blood cells from patients are capable of initiating leukaemia in mice, and that differentiated cells derived from the AML cannot initiate a tumour⁴³.

Stem-cell number and growth rate are maintained by multiple signals received from the microenvironment or niche^{44–46}, and recent work indicates that, in at least some cases, the signals involved in embryonic patterning contribute to the signals defining the adult stem-cell niche. Such a role is perhaps clearest for Hh in maintenance of stem-cell number in the somatic epithelial component of the *Drosophila* ovary. Stem cells giving rise to follicle cells of developing egg chambers are lost in the absence of Hh signal or upon loss of positively acting pathway components. In contrast, the number of stem cells roughly doubles upon constitutive activation of the Hh pathway⁴⁵. Less systematic evidence of a similar role for Hh and Wnt pathways is emerging in mammals and will be reviewed below in the context of some of the tumours associated with these pathways.

The Wnt pathway in colon cancer

Analysis of adenoma formation in patients and mice heterozygous for the *APC* gene has increased our understanding of the early steps of colon tumorigenesis (reviewed in refs 3, 47). In a normal colon crypt, all cells are thought to be derived from 4–6 epithelial stem cells that reside in the bottom of the crypt⁴⁸. In the colon, cells derived from a single crypt form a hexagonally shaped cuff of surface epithelium, whereas in the small intestine, multiple crypts supply cells to the epithelium of a single villus. The surface area of the intestinal epithelium seems to be maintained by the balance between the number of stem cells giving rise to newly differentiated cells and the number of

differentiated cells that die or that are shed. The conversion of all cells in a human crypt to a radiation-induced mutant phenotype occurs in approximately 1 year⁴⁹. This time is significantly longer than the two-day lifetime of differentiated enterocytes, and appears to represent the time it takes for a single mutant stem cell to take over the entire niche by the process of stem-cell competition.

Genetic studies in the mouse indicate that the transcription factor Tcf4 is required for maintenance of the stem-cell compartment in the small intestine⁵⁰, which suggests a role for the Wnt pathway in stem-cell maintenance (although expression of a Wnt ligand in this tissue has not yet been described). No phenotype of Tcf4 mutation is observed in the colon, where another member of the Tcf family, Tcf3, is also expressed⁵⁰. Based in part on the phenotype of Tcf4 loss of function, constitutively activated Tcf-mediated transcription in the crypt cells has been suggested to expand the stem-cell population³. A stem-cell origin for colon adenocarcinoma is consistent with the presence of multiple differentiated cell types in premalignant lesions⁵¹. These Wnt-directed transcriptional events in the intestine apparently provide a modestly expanded population of stem cell-like cells, which alone is insufficient to cause malignancy. But such cells would persist and would possess both unlimited replicative potential and partial self-sufficiency with respect to niche signals. At the same time, these premalignant cells would constitute targets for secondary mutations (such as those affecting the TGF- β response) that enhance growth and invasiveness, thus resulting in progression to malignancy³.

The Hedgehog pathway in tumorigenesis

Evidence regarding the role of Shh in medulloblastoma has emerged from recent analysis of the normal role of Shh in development of the cerebellum. During late embryogenesis, cells committed to the cerebellar granule-cell lineage migrate from the rhombic lip onto the roof of the cerebellar anlage to form the external germinal layer (EGL). In the EGL, these cells differentiate to granule-cell precursors and proliferate extensively. Subsequently, the cells migrate inward and terminally differentiate to form the internal granule layer of the cerebellum⁵². The proliferation of granule-cell precursors depends on Shh-secreting Purkinje cells. *In vitro*, Shh strongly inhibits terminal differentiation of mouse granule-cell precursors, and maintains their high proliferation rate^{53,54}. However, Shh induction of EGL cell proliferation after 9 h of incubation in serum-free media is markedly reduced⁵⁴. In the continued presence of Shh, most EGL-derived granule-cell precursors stop dividing at a time corresponding to postembryonic day 14, whatever the time of culture initiation⁵⁵. During the culture period, the proliferating granule-cell precursors accumulate p27 cyclin-dependent kinase inhibitor protein. A small subset of the EGL cells, however, continues to divide and fails to accumulate p27 (ref. 55).

These results are consistent with a model in which Shh inhibits terminal differentiation, allowing the granule-cell precursors to proliferate up to postembryonic day 14, an intrinsic time limit related perhaps to the accumulation of p27 (see above). In contrast, lack of Shh signal results in immediate loss of proliferative capacity. In addition, the EGL seems to contain a small percentage of cells that are not subject to the replicative limit, as indicated by their lack of p27 accumulation. In this context, it is interesting to note that some cells in the EGL continue to express nestin, a neuroectodermal stem-cell marker⁵². These stem cell-like cells may correspond to the cells that do not accumulate p27, and Shh might maintain these cells in a state that permits unlimited proliferative capacity. Medulloblastoma is thought to originate from the EGL granule-cell precursors^{53,56,57} and thus may be caused by Hh pathway-induced expansion of neural progenitors. This is consistent with the observation that desmoplastic or nodular medulloblastoma, which has been linked to mutations in *PTCH*⁵⁶, is characterized by nodules of differentiated cells expressing p27, surrounded by more primitive proliferating cells that do not stain for p27 (ref. 58). The route by which Shh may regulate medulloblastoma cell proliferation is through suppression of

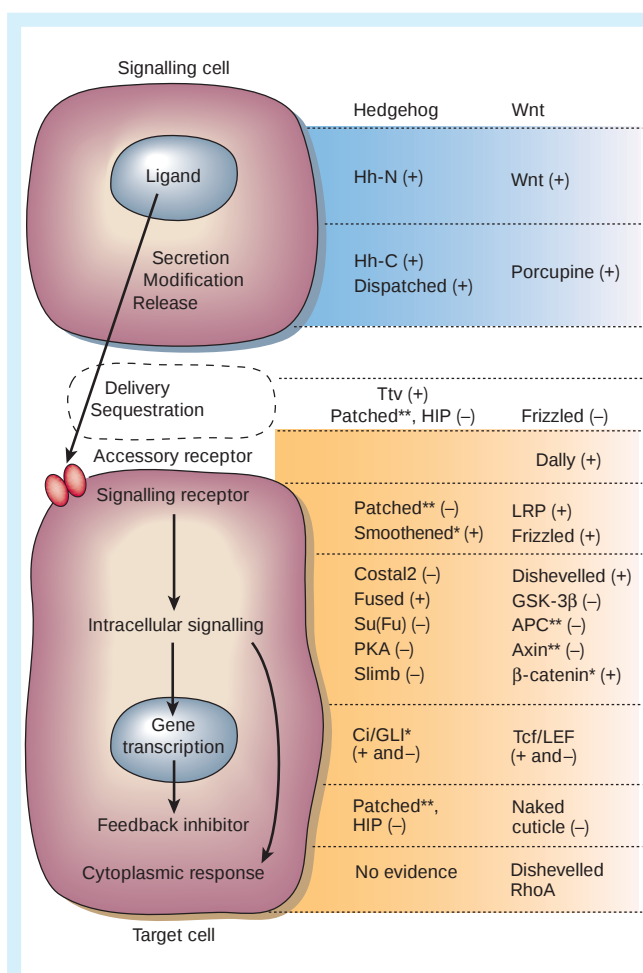


Figure 2 General characteristics and components of the Hh and Wnt signalling pathways. Functional characteristics of molecules acting in these pathways are listed on the left. In the producing (signalling) cell, multiple factors affect ligand secretion, modification or release. If the receiving cell is not adjacent to the producing cell, factors in cells in between affect ligand transport and sequestration. In the receiving (target) cell, the signal is interpreted by two classes of receptors, the accessory receptor and the signalling receptor, which bind to ligand and mediate transduction of the signal into the cytoplasm. In the cytoplasm, factors propagate the signal to effect a transcriptional response and a cytoplasmic response. Transcriptional activation leads to production of feedback inhibitors that limit pathway activity and decrease stochastic variation²¹. Individual components of the pathways are listed on the right (some have multiple isoforms; see refs 1–7). With respect to the transcriptional response, '+' and '–' indicate a positive and negative action, respectively. Components of the pathways found to be mutated or amplified in human cancer are denoted by asterisks (*, oncogene; **, tumour suppressor). Oncogenic mutations in these components lead to ligand-independent transcriptional activation. (Source: refs 1–7, 12, 13, 24, 32–34, 36, 47).

retinoblastoma (Rb) function, which is hyperphosphorylated (and thus inactivated) in EGL cells treated with Shh⁵⁴. Consistent with this possibility, medulloblastomas also arise in mice as a result of loss of p53 and Rb function⁵⁷.

Some clues have also emerged on how activation of the Hh pathway leads to formation of BCC, the most common type of cancer in Caucasians. Although classified as a malignant tumour, BCC is rarely metastatic and apparently arises without a pre-malignant lesion⁴¹. These observations indicate that initiation of BCC requires a relatively limited number of 'hits', and that perhaps a single lesion leading to constitutive activation of Gli-mediated transcription suffices³⁶. The origin of BCC is somewhat controversial⁴¹, but is likely to arise from an undifferentiated epithelial cell in the hair follicle⁵⁹. Hair follicle morphogenesis is regulated by Shh⁶⁰, and can be stimulated by

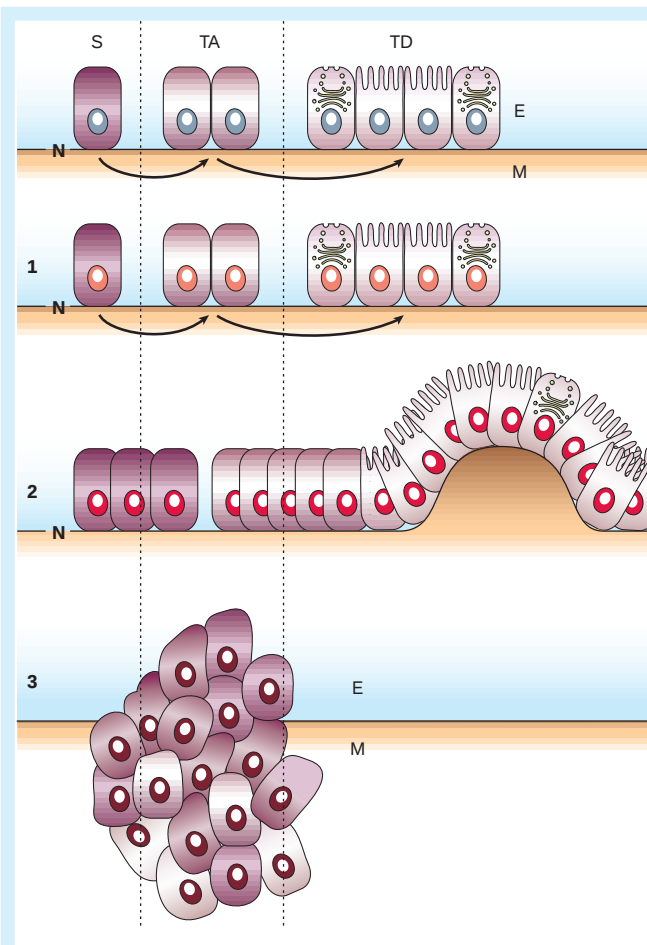


Figure 3 A model for tumorigenesis. The analysis of developmental pathways in cancer is consistent with the following model of tumorigenesis. Tumour initiation involves acquisition of morphologically silent mutations in a normal stem cell (stage 1, S). Such mutations may not directly confer a growth advantage, as they would affect 'caretaker' genes involved in maintenance of genome integrity⁴⁷ or single alleles of tumour suppressors such as *PTCH* or *APC* whose complete loss would directly contribute to the malignant phenotype. Genes involved in progression may also be mutated at this stage, but without a gross phenotypic alteration. These initial events are followed by a mutation leading to expansion of aberrant stem cells (stage 2), which give rise to transit amplifying cells (TA) that differentiate towards all or a subset of the normal terminally differentiated (TD) cell types (pre-malignant lesion). An increase in stem-cell number leads to disruption of spatial organization of cell proliferation, and proliferating cells are found in regions where no divisions normally occur. At this stage, the cells remain in the epithelial compartment (E). Subsequent mutations (stage 3) and clonal selection lead to more rapid proliferation, an increase in self-renewing divisions and a consequent decrease in differentiation. Further mutations (stage 3) also allow the cells to invade the mesenchyme (M) and metastasize. (See also refs 3, 41–43, 51, 67).

transient mis-expression of Shh⁶¹. One possibility is that BCC results from Shh pathway-dependent expansion of follicular stem cell-like cells⁴⁵. Consistent with this possibility, BCC is produced by overexpression of Gli1 in mice using the promoter for the keratin-5 gene, which is expressed in undifferentiated cells, including the hair follicle stem-cell compartment^{36,46}.

Implications and future directions

Current cancer therapies often engender severe toxicity because of their general effects on all rapidly dividing cells. Identification of candidate targets for more specific mechanism-based cancer therapy might use gene chip-based technologies, which could reveal signature patterns of transcriptional output characteristic of activated Wnt or Hh pathways. But a critical issue in mechanism-based

therapy⁶² is whether adults could tolerate inhibition of the Wnt and Hh pathways. This would be the case if tumorigenesis represented aberrant activation of pathways that normally are required only during embryonic development. However, emerging evidence suggests that these pathways control patterning and growth in self-renewing adult tissues by regulating the stem-cell compartment (Fig. 1). Thus, pharmacological inhibition of these pathways in the worst case might result in severe toxicity due to loss of normal stem-cell compartments. However, teratological evidence indicates that transient inhibition of the Shh pathway, sufficient to cause extreme malformations in embryos, is tolerated by pregnant females of various animal species⁶. Further research will be needed to determine whether continuous pathway activity is required in normal and tumour tissues, and whether these requirements differ sufficiently as to allow therapeutic intervention.

Even if pathway inhibition is prohibited by normal physiological requirements, other mechanism-based approaches that exploit aberrant pathway activation might be feasible. It has been proposed that malignancy is determined in all tissues by mis-regulation of a common set of cellular functions that control growth by affecting cell proliferation, apoptosis, invasion and angiogenesis⁴⁰. This hypothesis is supported by the demonstration that multiple types of normal human cells can be made tumorigenic by expression of a defined set of viral and cellular proteins^{63,64}. The high frequency of mutations in the Hh and Wnt pathways in specific tumour types suggests that the activated transcriptional output of these pathways efficiently mis-regulates multiple growth-regulatory functions. But the restricted range of tumour types associated with mutations in the Wnt and Hh pathways further argues that initiation of tumorigenesis requires the presence of other factors characteristic of the differentiated state of the target cell. Therapeutic agents for treatment of such tumours thus might target not only pathway components, but also other critical transcriptional targets of the Wnt and Hh pathways, or proteins that cooperate with them to mis-regulate growth. □

- Wodarz, A. & Nusse, R. Mechanisms of Wnt signaling in development. *Annu. Rev. Cell. Dev. Biol.* **14**, 59–88 (1998).
- Polakis, P. Wnt Signaling and cancer. *Genes Dev.* **14**, 1837–1851 (2000).
- Bien, M. & Clevers, H. Linking colorectal cancer to Wnt signaling. *Cell* **103**, 311–320 (2000).
- Ingham, P. W. Transducing hedgehog: the story so far. *EMBO J.* **17**, 3505–3511 (1998).
- Goodrich, L. V. & Scott, M. P. Hedgehog and patched in neural development and disease. *Neuron* **21**, 1243–1257 (1998).
- Muenke, M. & Beachy, P. A. In *The Metabolic and Molecular Bases of Inherited Disease* (eds Scriver, C., Beaudet, A., Sly, W. & Valle, D.) 6203–6230 (McGraw-Hill, New York, 2001).
- McMahon, A. P. More surprises in the hedgehog signaling pathway. *Cell* **100**, 185–188 (2000).
- Lee, J. J., von Kessler, D. P., Parks, S. & Beachy, P. A. Secretion and localized transcription suggest a role in positional signaling for products of the segmentation gene *hedgehog*. *Cell* **71**, 33–50 (1992).
- Kalderon, D. Transducing the Hedgehog signal. *Cell* **103**, 371–374 (2000).
- Porter, J. A., Young, K. E. & Beachy, P. A. Cholesterol modification of Hedgehog signaling proteins in animal development. *Science* **274**, 255–259 (1996).
- Pepinsky, R. B. et al. Identification of a palmitic acid-modified form of human Sonic hedgehog. *J. Biol. Chem.* **273**, 14037–14045 (1998).
- Baeg, G.-H. & Perrimon, N. Functional binding of secreted molecules to heparan sulfate proteoglycans in *Drosophila*. *Curr. Opin. Cell Biol.* **12**, 575–580 (2000).
- Burke, R. et al. Dispatched, a novel sterol sensing domain protein dedicated to the release of cholesterol-modified Hedgehog from signaling cells. *Cell* **99**, 803–815 (1999).
- Davies, J. P., Chen, F. W. & Ioannou, Y. A. Transmembrane molecular pump activity of Niemann-Pick C1 protein. *Science* **290**, 2295–2298 (2000).
- Stone, D. M. et al. The tumor-suppressor gene *patched* encodes a candidate receptor for Sonic hedgehog. *Nature* **384**, 129–134 (1996).
- Denef, N., Neubuser, D., Perez, L. & Cohen, S. M. Hedgehog induces opposite changes in turnover and subcellular localization of patched and smoothened. *Cell* **102**, 521–531 (2000).
- Taipale, J. et al. Effects of oncogenic mutations in *Smoothened* and *Patched* can be reversed by cyclopamine. *Nature* **406**, 1005–1009 (2000).
- Aza-Blanc, P., Ramirez-Weber, F.-A., Laget, M.-P., Schwartz, C. & Kornberg, T. B. Proteolysis that is inhibited by Hedgehog targets cubitus interruptus protein to the nucleus and converts it to a repressor. *Cell* **89**, 1043–1053 (1997).
- Robbins, D. J. et al. Hedgehog elicits signal transduction by means of a large complex containing the kinesin-related protein Costal2. *Cell* **90**, 225–234 (1997).
- Chen, C. H. et al. Nuclear trafficking of Cubitus interruptus in the transcriptional regulation of Hedgehog target gene expression. *Cell* **98**, 305–316 (1999).
- Freeman, M. Feedback control of intercellular signalling in development. *Nature* **408**, 313–319 (2000).
- Park, H. L. et al. Mouse Gli1 mutants are viable but have defects in SHH signaling in combination with a Gli2 mutation. *Development* **127**, 1593–1605 (2000).
- Hoblmayer, B. et al. WNT signalling molecules act in axis formation in the diploblastic metazoan *Hydra*. *Nature* **407**, 186–189 (2000).

24. Pinson, K. I., Brennan, J., Monkley, S., Avery, B. J. & Skarnes, W. C. An LDL-receptor-related protein mediates Wnt signalling in mice. *Nature* **407**, 535–538 (2000).
25. Bhanot, P. *et al.* A new member of the Frizzled family from *Drosophila* functions as a Wingless receptor. *Nature* **382**, 225–230 (1996).
26. Behrens, J. *et al.* Functional interaction of β -catenin with the transcription factor LEF-1. *Nature* **382**, 638–642 (1996).
27. Nusse, R. & Varmus, H. E. Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome. *Cell* **31**, 99–109 (1982).
28. Morin, P. J. *et al.* Activation of β -catenin-Tcf signaling in colon cancer by mutations in β -catenin or APC. *Science* **275**, 1787–1790 (1997).
29. Lickert, H. *et al.* Wnt/ β -catenin signaling regulates the expression of the homeobox gene Cdx1 in embryonic intestine. *Development* **127**, 3805–3813 (2000).
30. He, T.-C. *et al.* Identification of c-MYC as a target of the APC pathway. *Science* **281**, 1509–1512 (1998).
31. Shih, I.-M., Yu, J., He, T.-C., Vogelstein, B. & Kinzler, K. W. The β -catenin binding domain of adenomatous polyposis coli is sufficient for tumor suppression. *Cancer Res.* **60**, 1671–1676 (2000).
32. Johnson, R. L. *et al.* Human homolog of *patched*, a candidate gene for the basal cell nevus syndrome. *Science* **272**, 1668–1671 (1996).
33. Hahn, H. *et al.* Mutations of the human homolog of *Drosophila patched* in the nevoid basal cell carcinoma syndrome. *Cell* **85**, 841–851 (1996).
34. Xie, J. *et al.* Activating *Smoothed* mutations in sporadic basal-cell carcinoma. *Nature* **391**, 90–92 (1998).
35. Lam, C.-W. *et al.* A frequent activated *smoothed* mutation in sporadic basal cell carcinomas. *Oncogene* **18**, 833–836 (1999).
36. Nilsson, M. *et al.* Induction of basal cell carcinomas and trichoepitheliomas in mice overexpressing Gli-1. *Proc. Natl Acad. Sci. USA* **97**, 3438–3443 (2000).
37. Bray, S. A Notch affair. *Cell* **93**, 499–503 (1998).
38. Massagué, J., Blain, S. W. & Lo, R. S. TGF β signaling in growth control, cancer, and heritable disorders. *Cell* **103**, 295–309 (2000).
39. Taipale, J., Saharinen, J. & Keski-Oja, J. Extracellular matrix-associated transforming growth factor- β activation and roles in cancer cell growth and invasion. *Adv. Cancer Res.* **75**, 87–134 (1998).
40. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
41. Brash, D. E. & Pontén, J. Skin precancer. *Cancer Surv.* **32**, 69–113 (1998).
42. Jonason, A. S. *et al.* Frequent clones of p53-mutated keratinocytes in normal human skin. *Proc. Natl Acad. Sci. USA* **93**, 14025–14029 (1996).
43. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Med.* **3**, 730–737 (1997).
44. Xie, T. & Spradling, A. C. A Niche maintaining germ line stem cells in the *Drosophila* ovary. *Science* **290**, 328–330 (2000).
45. Zhang, Y. & Kalderon, D. Hedgehog acts as a somatic stem cell factor in the *Drosophila* ovary. *Nature* **410**, 599–604 (2001).
46. Fuchs, E. & Segre, J. A. Stem cells: a new lease on life. *Cell* **100**, 143–155 (2000).
47. Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170 (1996).
48. Booth, C. & Potten, C. Gut instincts: thoughts on intestinal epithelial stem cells. *J. Clin. Invest.* **105**, 1493–1499 (2000).
49. Campbell, F. *et al.* Post-irradiation somatic mutation and clonal stabilisation time in the human colon. *Gut* **39**, 569–573 (1996).
50. Korinek, V. *et al.* Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Nature Genet.* **19**, 379–383 (1998).
51. Moser, A. R., Dove, W. F., Roth, K. A. & Gordon, J. I. The Min (multiple intestinal neoplasia) mutation: its effects on gut epithelial cell differentiation and interaction with a modifier system. *J. Cell Biol.* **116**, 1517–1526 (1992).
52. Alder, J., Cho, N. K. & Hatten, M. E. Embryonic precursor cells from the rhombic lip are specified to a cerebellar granule neuron identity. *Neuron* **17**, 389–399 (1996).
53. Wechsler-Reya, R. J. & Scott, M. P. The developmental biology of brain tumors. *Annu. Rev. Neurosci.* **24**, 385–428 (2001).
54. Kenney, A. M. & Rowitch, D. H. Sonic hedgehog promotes G1 cyclin expression and sustained cell cycle progression in mammalian neuronal precursors. *Mol. Cell Biol.* **20**, 9055–9067 (2000).
55. Miyazawa, K. *et al.* A Role for p27/Kip1 in the control of cerebellar granule cell precursor proliferation. *J. Neurosci.* **20**, 5756–5763 (2000).
56. Pietsch, T. *et al.* Medulloblastomas of the desmoplastic variant carry mutations of the human homologue of *Drosophila patched*. *Cancer Res.* **57**, 2085–2088 (1997).
57. Marino, S., Vooijs, M., van der Gulden, H., Jonkers, J. & Berns, A. Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of Rb in the external granular layer of the cerebellum. *Genes Dev.* **14**, 994–1004 (2000).
58. Schiffer, D. *et al.* Cell-cycle inhibitor p27/Kip-1 expression in non-astrocytic and non-oligodendrocytic human nervous system tumors. *Neurosci. Lett.* **264**, 29–32 (1999).
59. Krüger, K., Blume-Peytavi, U. & Orfanos, C. E. Basal cell carcinoma possibly originates from the outer root sheath and/or the bulge region of the vellus hair follicle. *Arch. Dermatol. Res.* **291**, 253–259 (1999).
60. St-Jacques, B. *et al.* Sonic hedgehog signaling is essential for hair development. *Curr. Biol.* **8**, 1058–1068 (1998).
61. Sato, N., Leopold, L. P. & Crystal, R. G. Induction of the hair growth phase in postnatal mice by localized transient expression of Sonic hedgehog. *J. Clin. Invest.* **104**, 855–864 (2000).
62. Gibbs, J. B. Mechanism-based target identification and drug discovery in cancer research. *Science* **287**, 1969–1973 (2000).
63. Hahn, W. C. *et al.* Creation of human tumor cells with defined genetic elements. *Nature* **400**, 464–468 (1999).
64. Elenbaas, B. *et al.* Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. *Genes Dev.* **15**, 50–65 (2001).
65. Conlon, I. & Raff, M. Size control in animal development. *Cell* **96**, 235 (1999).
66. Counter, C. M. *et al.* Dissociation among in vitro telomerase activity, telomere maintenance, and cellular immortalization. *Proc. Natl Acad. Sci. USA* **95**, 14723–14728 (1998).
67. Tsao, J.-L. *et al.* Tracing cell fates in human colorectal tumors from somatic microsatellite mutations: evidence for adenomas with stem cell architecture. *Am. J. Pathol.* **153**, 1189–1200 (1998).
68. Tejpar, S. *et al.* Predominance of β -catenin mutations and β -catenin dysregulation in sporadic aggressive fibromatosis (desmoid tumor). *Oncogene* **18**, 6615–6620 (1999).
69. Wei, Y. *et al.* Activation of β -catenin in epithelial and mesenchymal hepatoblastomas. *Oncogene* **19**, 498–504 (2000).
70. Sparks, A. B., Morin, P. J., Vogelstein, B. & Kinzler, K. W. Mutational analysis of the APC/ β -catenin/Tcf pathway in colorectal cancer. *Cancer Res.* **58**, 1130–1134 (1998).

Acknowledgements

We apologize to the authors of original work not cited. Due to editorial limits on references, we were largely limited to citing review articles and more recent original articles that have not been reviewed previously. We thank D. M. Berman for critical review of the manuscript. P.A.B. is an investigator of the Howard Hughes Medical Institute. This work was supported by a grant from the NIH.

Oncogenic kinase signalling

Peter Blume-Jensen* & Tony Hunter

The Salk Institute, Molecular and Cell Biology Laboratory, 10010 North Torrey Pines Road, La Jolla, California 92037, USA

(e-mail: blume@salk.edu; hunter@salk.edu)

*Current address: Serono Reproductive Biology Institute, 280 Pond Street, Randolph, Massachusetts 02368, USA

Protein-tyrosine kinases (PTKs) are important regulators of intracellular signal-transduction pathways mediating development and multicellular communication in metazoans. Their activity is normally tightly controlled and regulated. Perturbation of PTK signalling by mutations and other genetic alterations results in deregulated kinase activity and malignant transformation. The lipid kinase phosphoinositide 3-OH kinase (PI(3)K) and some of its downstream targets, such as the protein-serine/threonine kinases Akt and p70 S6 kinase (p70^{S6K}), are crucial effectors in oncogenic PTK signalling. This review emphasizes how oncogenic conversion of protein kinases results from perturbation of the normal autoinhibitory constraints on kinase activity and provides an update on our knowledge about the role of deregulated PI(3)K/Akt and mammalian target of rapamycin/p70^{S6K} signalling in human malignancies.

Deregulated (that is, autonomous) cell growth is the defining feature of all neoplasms, both benign and malignant. Malignant neoplasms have, in addition, the capacity to invade normal tissues and metastasize to and grow at distant body sites, the other main defining criterion of cancer. Deregulated cell growth occurs as a result of perturbed signal transduction defined, in its broadest sense, as all cellular signals that modulate or alter cellular behaviour or function^{1,2}. Consequently, cancers do not necessarily arise as a result of an increased rate of cell proliferation. Rather, it is the critical balance between the rate of cell-cycle progression (cell division) and cell growth (cell mass) on one hand, and programmed cell death (apoptosis) on the other, that is important³. During normal embryonic development and in adult life, signalling needs to be precisely coordinated and integrated at all times, and properly regulated differentiation signals are critical for preventing oncogenesis. The old dogma stating an inverse relation between cell-differentiation stage and (deregulated) cell proliferation (a malignant tumour tends to be more de-differentiated than its parent cell type) illustrates this important principle.

Certain classes of signalling proteins and pathways are targeted much more frequently by oncogenic mutations than others^{4,5}. Thus, molecules governing extracellular growth, differentiation and developmental signals, in particular, are often mutated in cancers. One illustrative example is provided by receptor protein-tyrosine kinases (RPTKs), a subclass of transmembrane-spanning receptors endowed with intrinsic, ligand-stimulatable PTK activity. When mutated or altered structurally, RPTKs can become potent oncoproteins, causing cellular transformation. Conversely, RPTK activity in resting, untransformed cells is normally tightly controlled. Recent studies have provided new insights regarding the structural bases for normal intramolecular control of PTKs, and shown that multiple 'layers' of autoinhibitory mechanisms operate as a safeguard against unwanted protein kinase activation. Hence, rather than looking at oncogenic mechanisms as 'activating events' or 'hits', whether due to mutations, overexpression or structural re-arrangements, it is more meaningful to see them as mechanisms causing primarily relief or obstruction of normal autoinhibitory and regulatory constraints. This review presents some examples of perturbed PTK signalling

mechanisms, as well as providing an update on recent insights regarding the role of PI(3)K and ribosomal p70^{S6K} in oncogenesis.

RPTK regulation by autoinhibition

The sequencing effort of the Human Genome Project has revealed that up to ~20% of the ~32,000 human coding genes encode proteins involved in signal transduction, including transmembrane receptors, G-protein subunits and signal-generating enzymes. Among these, are more than 520 protein kinases and 130 protein phosphatases, exerting tight and reversible control on protein phosphorylation. Both of these enzyme categories can be subdivided into tyrosine- or serine/threonine-specific, based on their catalytic specificity. In addition, some possess dual specificity for both tyrosine and serine/threonine, and a few members of the phosphatidylinositol kinase family also exhibit protein-serine/threonine kinase activity. At the time of writing there are >90 known PTK genes in the human genome; 58 encode transmembrane RPTKs distributed into 20 subfamilies, and 32 encode cytoplasmic, non-receptor PTKs in 10 subfamilies^{6,7} (Figs 1, 2). Of the ~30 tumour-suppressor genes and >100 dominant oncogenes known to date⁸, protein kinases, in particular PTKs, comprise a large fraction of the latter group (Table 1). PTKs are also the largest group of dominant oncogenes with structural homology. PTKs evolved to mediate aspects of multicellular communication and development; they are found only in metazoans, where they comprise ~0.3% of genes. Somatic mutations in this very small group of genes cause a significant fraction of human cancers, again emphasizing the inverse relationship between normal developmental regulation and oncogenesis.

The physiological regulation of RPTKs is key to understanding the mechanisms causing their oncogenic activation (see refs. 9,10 for recent reviews). Signalling by RPTKs requires ligand-induced receptor oligomerization, which results in tyrosine autophosphorylation of the receptor subunits¹¹. This both activates catalytic activity and generates phosphorylated tyrosine residues that mediate the specific binding of cytoplasmic signalling proteins containing Src homology-2 (SH2) and protein tyrosine-binding (PTB) domains. Crystal structures of the inactive forms of the core cytoplasmic kinase domains of the insulin receptor (InsR), fibroblast growth factor receptor-1

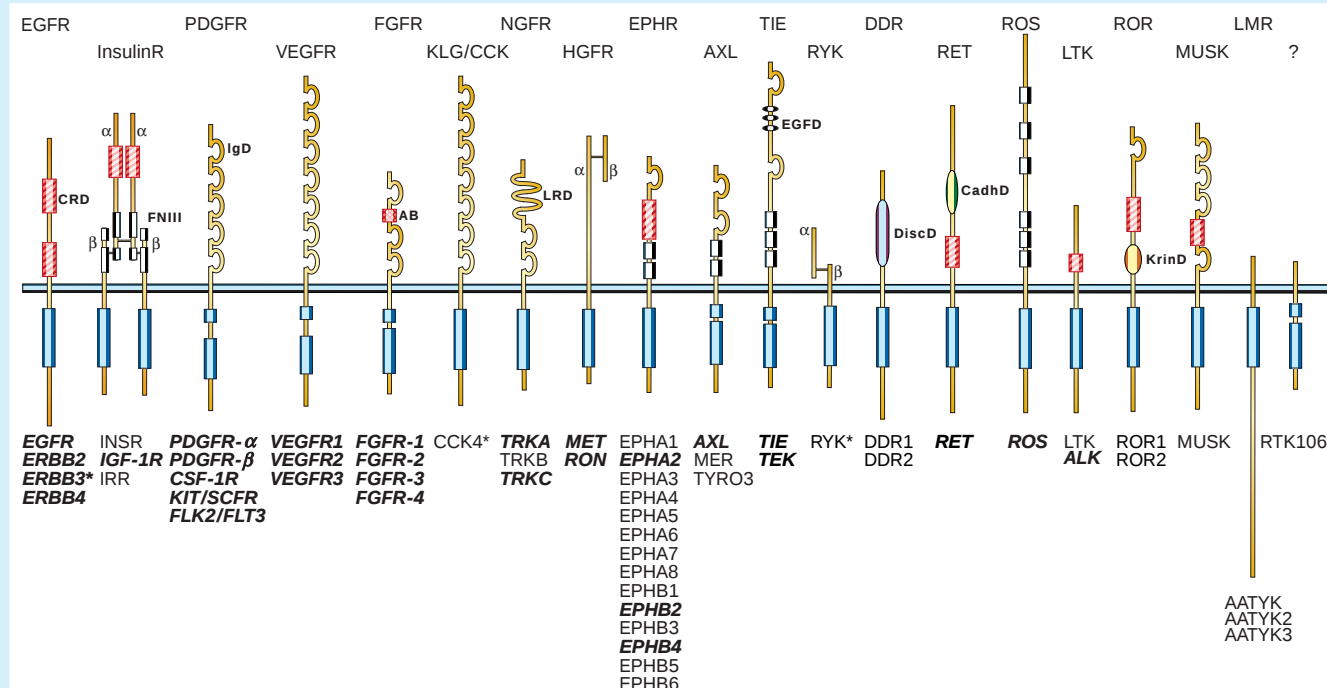


Figure 1 Human receptor protein-tyrosine kinases. The prototypic receptor for each family is indicated above the receptor, and the known members are listed below. Abbreviations of the prototypic receptors: EGFR, epidermal growth factor receptor; INSR, insulin receptor; PDGFR, platelet-derived growth factor receptor; VEGFR, vascular endothelial growth factor receptor; FGFR, fibroblast growth factor receptor; KLG/CCK, colon carcinoma kinase; NGFR, nerve growth factor receptor; HGFR, hepatocyte growth factor receptor; EphR, ephrin receptor; Axl, a Tyro3 PTK; TIE, tyrosine kinase receptor in endothelial cells; RYK, receptor related to tyrosine kinases; DDR, discoidin domain receptor; Ret, rearranged during transfection; ROS, RPTK expressed in some epithelial cell types; LTK, leukocyte tyrosine kinase; ROR, receptor orphan; MuSK, muscle-specific kinase; LMR, Lemur. Other abbreviations: AB, acidic box; CadhD, cadherin-like domain; CRD, cysteine-rich domain; DiscD, discoidin-like domain; EGFD, epidermal growth factor-like domain; FNIII, fibronectin type III-like domain; IgD, immunoglobulin-like domain; KrinD, kringle-like domain; LRD, leucine-rich domain. The symbols α and β denote distinct RPTK subunits. RPTK members in bold and italic type are implicated in human malignancies (see Table 1). An asterisk indicates that the member is devoid of intrinsic kinase activity.

(FGFR1) and Flk1 (vascular endothelial growth factor receptor-2 or VEGFR2) tyrosine kinases have provided a molecular understanding of the tight control of catalytic activity resulting from a *cis* inhibition/*trans* activation mechanism¹².

In both the unstimulated FGFR1 and VEGFR2, the activation loop occludes substrate tyrosine binding to the active site, whereas in the unstimulated INSR, Tyr1162 in the activation loop is bound in *cis* in the active site to prevent substrate access, and the beginning of the activation loop occludes ATP binding. In response to ligand stimulation of RPTKs, one or several of the tyrosine residues in the activation loop is phosphorylated *trans* by the dimeric receptor partner. This leads to repositioning of the activation loop away from the active site, allowing substrate (and in the case of the INSR, ATP) access. An equilibrium exists between inactive and active loop conformations of the unphosphorylated RPTKs in solution¹³. This equilibrium enables *cis* inhibition, yet allows phosphorylation in *trans* between ligand-induced receptor dimers. The importance of this is supported by several examples of oncogenic point mutations in the activation loop that cause constitutive RPTK activation. A simplified diagram of RPTK activation is presented in Fig. 3a.

Evidence indicates that RPTK dimerization per se is not always sufficient for kinase activation. There seems to be an additional requirement for ligand-induced conformational switches, ensuring that the catalytic domains are juxtaposed in a proper configuration to enable phosphorylation in *trans* between receptor subunits^{9,14}. Accordingly, other regions of RPTKs can have autoinhibitory functions in unstimulated cells. For instance, in Tek (TIE2), the carboxy-terminal tail partially occludes the substrate tyrosine-binding site¹⁵. Autophosphorylation of the tail region exposes its phosphotyrosine residues for substrate binding, as well as the substrate-binding site in the kinase. In addition, for some RPTKs, including platelet-derived

growth factor receptor (PDGFR), Kit/stem-cell factor receptor (SCFR), colony-stimulating factor-1 receptor (CSF1-R), ephrin receptor (EphR) and INSR, the juxtamembrane region has been implicated in autoinhibition. Hence, autophosphorylation of one or two homologous juxtamembrane tyrosine residues in several of these receptors is required for full kinase activation, and mutation to phenylalanine significantly reduces kinase activation (see ref. 16 and references therein).

Autophosphorylation of these residues seems to serve a dual function: it relieves the inhibitory conformation enabling full kinase activation and at the same time creates binding sites for numerous SH2-containing signalling molecules, such as Src, RasGAP, SHP-1, SHEP-1 and PI(3)K. Consistent with a repressive function of the juxtamembrane region, substitution of a Val residue just amino-terminal to the regulatory tyrosine residues in the PDGFR- β results in constitutive kinase activation¹⁷. In addition, numerous oncogenic mutations in human Kit are located either N- or C-terminal to the two tyrosines, and internal gene duplications in the juxtamembrane region of Flk2/Flt3 result in constitutive kinase activity^{18,19}. The crystal structure of the EphB2 kinase domain has revealed a possible mechanism for the inhibition by its juxtamembrane region. The unphosphorylated juxtamembrane region impinges on the C α helix in the N-terminal kinase lobe and other regions of the kinase, resulting in catalytic repression. The structure also suggests how phosphorylation relieves repression by causing dissociation of the juxtamembrane region (L. Groot, B. Baskin, T. Pawson & F. Sicheri, personal communication).

Deregulation of RPTK by relief of restraints

How do these insights apply to cellular transformation and cancer? In principle, for all PTKs involved in cancer, oncogenic deregulation results from relief or perturbation of one or several of the

auto-control mechanisms that ensure the normal repression of catalytic domains. A little more than half (31) of the known RPTKs have been repeatedly found in either mutated or overexpressed forms associated with (human) malignancies, including sporadic cases (see Table 1).

There are four main principles for oncogenic transformation by PTKs. First, retroviral transduction of a proto-oncogene corresponding to a PTK concomitant with deregulating structural changes is a common transforming mechanism in rodents and chicken. Second, genomic re-arrangements, such as chromosomal translocations, can result in oncogenic fusion proteins that include (minimally) a PTK catalytic domain and an unrelated protein that provides a dimerization function. Third, gain-of-function (GOF) mutations or small deletions in RPTKs and cytoplasmic PTKs are associated with several malignancies. Finally, PTK overexpression resulting from gene amplification is associated with several common human cancers. In general, the transforming effect can be ascribed to enhanced or constitutive kinase activity with quantitatively or qualitatively altered downstream signalling.

RPTK overexpression leads to constitutive kinase activation by increasing the concentration of dimers. Important examples are the Neu/ErbB2 and epidermal growth factor receptor (EGFR), which are often amplified in breast and lung carcinomas (Table 1). ErbB2 signalling is inhibited by binding of the monoclonal antibody herceptin, which is being used in the treatment of ErbB2-positive breast cancers. A selective small-molecule EGFR-tyrosine kinase inhibitor, ZD1839 ('Iressa'), is in late phase trials for advanced non-small-cell lung cancer. Cancers due to chimaeric RPTK domains also involve constitutive kinase activation, which depends on the oligomerization domain(s) of the N-terminal fusion partner. Enforced dimer formation juxtaposes the catalytic domains in an optimal orientation for *trans* phosphorylation, probably very similar to a ligand-induced receptor dimer.

Among the more than 30 RPTKs implicated in human cancer, some, such as ErbB2 and EGFR that are amplified, mutated and/or overexpressed in prevalent cancers (Table 1), have been extensively reviewed (see, for example, ref. 20). Here, we will restrict our discussion to the Ret/glia-derived neurotrophic factor receptor

(GDNFR) and Kit/SCFR RPTKs, which illustrate a variety of mechanisms for RPTK-induced oncogenic transformation.

Ret/GDNFR

Ret is required for development of the kidneys and enteric system and for neuronal differentiation and survival. It is part of a multicomponent receptor for the glial-derived neurotrophic factor (GDNF) family of neurotrophins, which include neurturin (NTN), artemin (ART) and persephin (PSP). In response to ligand, Ret is activated by heterodimer formation with one of four structurally related glycosylphosphatidylinositol (GPI)-linked cell-surface receptors, GFR- α 1–4. At least eight common somatic rearrangements result in fusions between the N terminus of various proteins and the PTK domain of Ret (see Table 1 and ref. 21). This leads to subsequent GDNF- and NTN-independent kinase activation caused by constitutive dimerization of the fusion proteins, resulting in papillary thyroid carcinomas (PTCs). Somatic Ret GOF mutations are found in some sporadic tumours, while germline Ret GOF mutations are involved in three familial tumour syndromes — multiple endocrine neoplasia 2A (MEN2A), MEN2B and familial medullary thyroid carcinoma (FMTC; see Table 1).

All the identified oncogenic mutations cluster in extracellular-domain exons 10, 11 and 13–16 of the PTK domain. Almost 100% of patients with MEN2A and FMTC have mutations in one of the five conserved cysteines in the extracellular domain of Ret, causing formation of intermolecular disulphide bonds between Ret molecules, and constitutive dimerization and activation²². In contrast, MEN2B is due to a recurring Met918Thr mutation (a methionine-to-threonine substitution at codon 918), which activates by a different mechanism. Met918 corresponds to a highly conserved Met just upstream of the Ala-Pro-Glu motif (PTK subdomain VIII) in the substrate-binding pocket of RPTKs, whereas a threonine residue at this position is typical for cytoplasmic PTKs, like c-Src. This replacement increases the kinase activity of Ret without constitutive dimer formation²², most likely because it enables substrate access without prior autophosphorylation of the activation loop. In addition, it alters the substrate specificity of the Ret/MEN2B receptor towards peptide substrates that are optimal for Src and Abl^{23,24}. The altered specificity leads to autophosphorylation of Ret on novel tyrosine

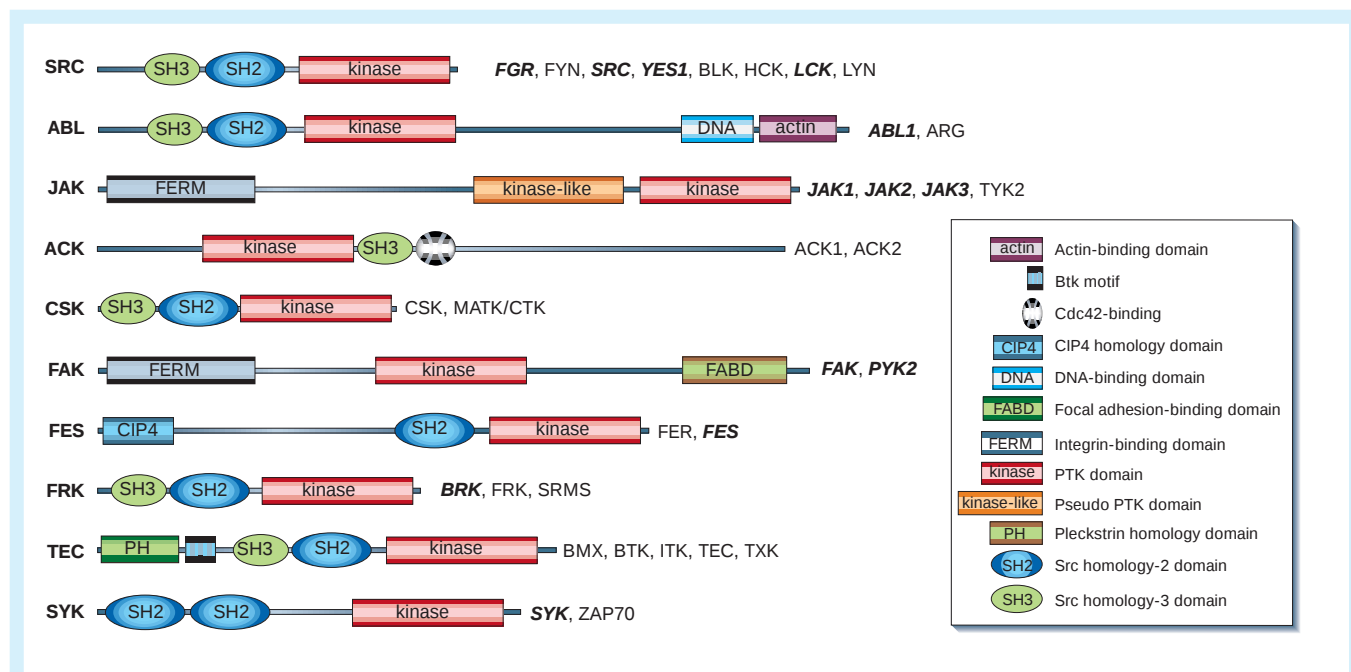
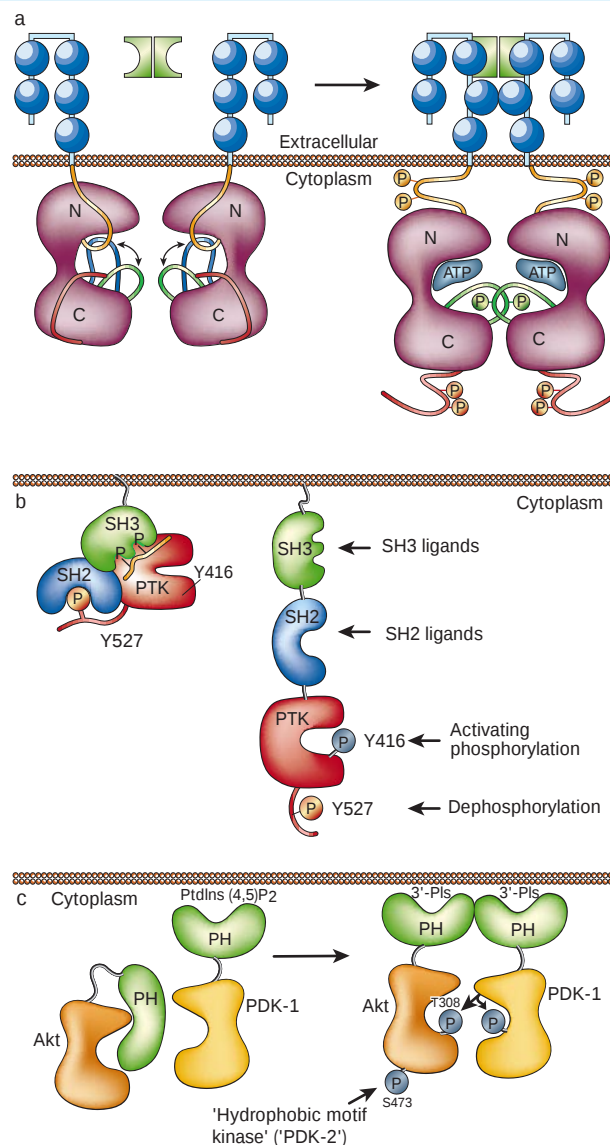


Figure 3 Protein kinase activation mechanisms. **a**, RPTK activation. Left: RPTK kinase activity is tightly repressed in the unstimulated state. The activation and catalytic loops exist in an equilibrium between a substrate-precluding (blue) and substrate-accessible (green) conformation. In addition, the juxtamembrane region (orange) and C-terminal region (red) might interfere with the conformation of the N-terminal kinase lobe ('N') and/or substrate access. Right: ligand-induced receptor dimerization and tyrosine autophosphorylation result in relief of the inhibitory constraints exerted by the activation loop, and the juxtamembrane and C-terminal regions. **b**, c-Src activation. Left: c-Src kinase activity is tightly repressed in the unstimulated state. The SH2 domain interacts with phospho-Tyr 527 in the C terminus and the SH3 domain with the polyproline type II helix in the linker region between the SH2 and kinase domain. This causes misalignment of residues that are critical for kinase activity. Right: binding of ligands to the SH2 or SH3 domain and/or dephosphorylation of phospho-Tyr 527 by PTPs relieves the inhibitory constraints on the kinase. **c**, Akt activation. Left: it is thought that the N-terminal PH domain precludes kinase access to and phosphorylation of the activation-loop Thr 308 by PDK-1. Right: PI(3)K activation results in production of $\text{PtdIns}(3,4,5)\text{P}_3$ and $\text{PtdIns}(3,4)\text{P}_2$, which recruits Akt to the membrane by binding to its PH domain. This exposes Thr 308 for phosphorylation by PDK-1, which is already located at the membrane. An unidentified PDK-2 kinase phosphorylates Ser 473 in the C terminus, which leads to full Akt activation. See text for details.



residues and tyrosine phosphorylation of substrates that are not phosphorylated by the activated wild-type Ret. However, there is also increased activation of PI(3)K, which might be crucial for transformation²⁵.

A recently developed *in vivo* model for MEN2B, based on introduction of the Met918Thr homologous mutation into the germline of mice, will enable studies of different signal-transduction pathways from Ret/MEN2B involved in tumorigenesis²⁶. Besides the Met918Thr mutation, a number of other recurrent point mutations have been found in Ret in MEN2A, MEN2B and FMTC. In all cases, the activating mutations occur in highly conserved regions of the Ret PTK domain that are normally involved in kinase repression in the inactive receptor. Different tyrosine residues are likely to be autophosphorylated in the different Ret mutants, resulting in binding and activation of different signalling molecules²⁷. Based on mutagenesis studies, Grb2 and Shc acting upstream of the classical mitogenic Ras-Raf-ERK (for extracellular signal-regulated protein kinase) cascade and PI(3)K seem important for the transforming effects.

Kit/SCFR

The Kit/SCFR provided the first example of naturally occurring, germline loss-of-function (LOF) point mutations in a mammalian RPTK, and the resulting phenotypes established the importance of

this receptor for normal haematopoiesis and mast-cell development, melanogenesis, gametogenesis and development of interstitial cells of Cajal^{28,29}. Recently, more than 30 GOF mutations, either single amino-acid changes or deletions of a few amino acids, have been identified in the Kit/SCFR, and they are associated with several highly malignant tumours in humans (reviewed in ref. 30). The mutations tend to cluster in two regions. Those in exon 11 contained in the juxtamembrane region are associated with gastrointestinal stromal tumours, whereas recurrent exon 17-mutations of Asp816 to either Val or His in the second half of the kinase domain are associated with mast-cell/myeloid leukaemias and seminomas/dysgerminomas, respectively.

The transforming mechanism for both of the main types of mutation involves dimer formation resulting in constitutive ligand-independent kinase activation. In most patients, tumours are heterozygous for the mutant form of *c-kit*, which indicates a dominant-positive phenotype. This is consistent with constitutively active heterodimers formed between the mutant and wild-type Kit receptors. The mutations in the juxtamembrane region cluster around the two autophosphorylation sites, Tyr 568 and 570, involved in binding c-Src and SHP-1 (an SH2 domain-containing protein-tyrosine phosphatase (PTP) expressed in haematopoietic cells)^{31,32}, and the presence of mutant Kit implies a worse prognosis³³. The

juxtamembrane mutations probably relieve the repressive effect of this domain on PTK activity, allowing limited autophosphorylation of the kinase domain, which results in additional activating conformational changes enabling full kinase activation¹³. It is of note that the frequently mutated Asp816 in Kit is a highly conserved residue just C-terminal to the conserved Asp-Phe-Gly motif in the activation loop of protein kinases. Mutation of the corresponding residue in Met and Ret results in papillary renal and thyroid carcinomas, respectively³⁴. This mutation seems to shift the equilibrium of the activation loop in unstimulated RPTKs towards the active conformation¹³.

Among the many Kit-induced signalling pathways, PI(3)K is particularly important and seems crucial for oncogenic transformation by GOF point-mutant Kit receptors^{35,36}. SHP-1 might also be an important target for inactivation in Kit-induced oncogenesis. SHP-1 binds to the juxtamembrane tyrosine autophosphorylation sites in Kit and CSF-1R^{31,37}, resulting in phosphatase activation and direct receptor dephosphorylation. The importance of SHP-1 in negative regulation of Kit and CSF-1R signalling is supported by the phenotype of *Mothaten* (*Me*)-mutant mice, which have LOF mutations of SHP-1, and the partial phenotypic rescue obtained by crossing them to *Dominant white-spotting W*-mutant mice, which have naturally occurring LOF mutations in Kit³⁸. Asp816Val-Kit causes enhanced degradation of SHP-1 through the proteasome pathway³⁹, and alternative transcripts causing LOF mutations or truncations of SHP-1 are frequent in primary Kit-expressing tumours from leukaemic patients⁴⁰. Thus, SHP-1 might be a tumour suppressor for Kit-induced malignancies. Clinical trials with the small-molecule Bcr-Abl PTK inhibitor STI 571 (ref. 41), which also inhibits Kit/SCFR and PDGFR- β ⁴², have been initiated for Kit-positive gastrointestinal stromal tumours. The small-molecule PTK inhibitors SU5416 and SU6668, originally developed for VEGFR as angiogenesis inhibitors, also inhibit Kit and are entering early trials this year for Kit-positive acute myeloid leukaemias.

Besides regulation by PTPs, other mechanisms for inhibition of RPTK signalling include ligand-induced receptor endocytosis, regulation of negative feedback loops, and heterodimerization with kinase-inactive RPTKs. Perturbation of these inhibitory mechanisms might result in RPTK-induced malignancies or sensitize cells for oncogenic transformation under some circumstances.

Cytoplasmic protein-tyrosine kinases

Given their importance in receptor signalling pathways, it might come as a surprise that of the 32 known cytoplasmic PTKs (Fig. 2), less than half have been implicated convincingly in human cancer (Table 1). The bias comes from the fact that most of the initial studies were on the viral counterparts, and aimed at elucidating their transforming mechanisms.

c-Src

c-Src was the first cellular homologue of a viral oncoprotein to be discovered⁴³; it is important for mitogenic signalling from many RPTKs, and has been implicated in a variety of cancers (reviewed in ref. 44). Just as for RPTKs, there is normally tight control of c-Src kinase activity through intramolecular interactions. In inactive c-Src, a C-terminal tyrosine residue (527 in mouse, 530 in human c-Src), lacking in deregulated v-Src, is phosphorylated and interacts with the SH2 domain, while the c-Src SH3 domain interacts with the linker region between the SH2 domain and the N-terminal kinase lobe. The SH2 and SH3 intramolecular interactions repress kinase activity by displacing the α helix in the N-terminal lobe and by positioning the activation loop to block access to the active site. Accordingly, c-Src can be de-repressed not only upon dephosphorylation of phosphorylated Tyr 527, but also by binding through its SH2 domain to specific tyrosine autophosphorylation sites in ligand-stimulated RPTKs, resulting in SH2 displacement from phosphorylated Tyr 527, or by binding of the SH3 domain to Pro-X-X-Pro motifs in target proteins⁴⁵. This

results in autophosphorylation in *trans* of the conserved activation loop Tyr416 and stabilization of the active conformation. This activation mechanism is similar to that of RPTKs, in that the activating event (ligand binding and dimerization for RPTKs, and SH2- or SH3-domain ligand engagement for c-Src) results in removal of inhibitory constraints on the kinase domain (Fig. 3a, b).

Several PTPs are implicated in regulating c-Src through dephosphorylation of the C-terminal c-Src kinase (CSK) phosphorylation site, Tyr 527, including receptor-like PTP α , PTP λ and RPTP ϵ , and the cytoplasmic PTP1B, SHP-1 and SHP-2. Elevated activity or expression of several of these PTPs correlate with enhanced levels of c-Src kinase activity in a number of transformed cells⁴⁴. But the most direct demonstration that c-Src is involved in human cancer was the identification of a mutant c-Src with truncation of the C terminus, ending with Tyr 530 in human colon cancer⁴⁶. This mutant has deregulated kinase activity because the lack of residues C-terminal to the phosphorylated Tyr 530 prevents SH2 association and establishment of the inactive conformation. Again, oncogenic perturbation results from relief of the tightly controlled constraints on kinase activity. It is not clear which signalling pathways are important for c-Src transformation, but a dominant-negative mutant of signal transducer and activator of transcription (STAT)-3 blocks v-Src transformation and c-Myc induction, indicating that STATs might be involved.

c-Abl

The Philadelphia (Ph) chromosome provided the first example of a consistent chromosomal abnormality associated with a specific type of leukaemia⁴⁷. The t(9; 22) reciprocal translocation involves the non-receptor PTK c-Abl on chromosome 9 and a breakpoint cluster region on chromosome 22. A majority of patients with chronic myeloid leukaemia, and a significant fraction of Ph-positive patients with acute lymphocytic leukaemia have one of three different versions of this translocation (Table 1). The c-Abl protein is structurally complex, consisting of SH3, SH2, PTK, DNA-binding and actin-binding domains (Fig. 2), among others.

Recent evidence suggests that nuclear c-Abl has a role primarily in DNA damage-induced apoptosis, and that Bcr-Abl circumvents this function because it is retained in the cytoplasm⁴⁸. Hence, nuclear c-Abl is activated by ionizing radiation and certain cytostatic drugs, and the activation is dependent on ataxia telangiectasia-mutated (ATM), a protein-serine/threonine kinase belonging to a family of proteins possessing a phosphoinositide kinase-homology domain. Conversely, the tumour suppressor Rb, which binds to c-Abl in the G0 and G1 phases of the cell cycle and represses the tyrosine kinase function of c-Abl, prevents the DNA damage-induced activation of nuclear c-Abl, which is seen only after entry into S-phase, when c-Abl is released by Rb hyperphosphorylation. Consistently, Rb-deficient cells are more sensitive to DNA damage-induced cell death (reviewed in ref. 49).

The possible mediators of c-Abl-induced cell death include the transcription factors p53 and p73, and the stress-activated mitogen-activated protein kinase (MAPK) family members c-Jun N-terminal kinase (JNK/SAPK) and p38, but p73 seems to be the crucial target. Hence, p73 accumulates in both wild-type and p53-deficient cells in response to DNA damage in a c-Abl-dependent manner, and ectopic expression of p73 induces growth arrest and apoptosis, in part by inducing p53 target genes⁴⁹. The transcription factor E2F1, deregulated in many cancers and known to stabilize p53, directly transactivates p73, causing transcription of p53 target genes in a p53-independent manner, and apoptosis⁵⁰. E2F1 is released from Rb during G1 exit, and so the induction of p73 can occur only in early S phase. It will be interesting to see whether c-Abl-induced apoptosis via p73 is dependent on released E2F1, which would explain why Abl induces apoptosis only after Rb hyperphosphorylation in early S phase. The c-Abl protein also functions in the cytoplasm, where it is involved in PDGF-induced motility responses and cell adhesion⁵¹.

Bcr-Abl is localized exclusively in the cytoplasm of transformed cells by retention mechanisms that involve Abl kinase activity and Bcr

Table 1 Examples of dominant protein-tyrosine kinase oncogenes

PTK (proto-oncogene)	Viral oncogene* (viral oncoprotein)	Oncogenic alteration	Tumour/cancer types (only the most frequent, mainly human types are described)
EGFR/ErbB1 (c-erbB)	v-erbB from AEV (p68/74 ^{erbB})	v-erbB: Truncated EGFR PTK c-erbB: Overexpression (amplification) Extracellular domain deletions	v-ErbB: fibrosarcomas c-ErbB: mammary carcinoma, glioblastoma multiforme, ovarian, non-small-cell lung and other cancers
ErbB2/HER2/Neu		Overexpression (amplification) No recurrent human mutations (Val664Glu in rodents)	Mammary, ovarian, gastric, non-small-cell lung and colon cancer
ErbB3/HER3		Overexpression; constitutive tyrosine phosphorylation (heterodimer with ErbB2)	Mammary carcinoma
ErbB4/HER4		Overexpression	Mammary carcinoma, granulosa cell tumours
IGF-1R		Overexpression (expression required for <i>in vitro</i> transformation by many oncogenes and DNA viruses)	Cervical and other carcinomas, sarcomas
PDGFR- α		Overexpression (amplification)	Glioma, glioblastoma, ovarian carcinoma
PDGFR- β		Tel-PDGR- β (t(5; 12) translocation fusing Ets-like Tel with PDGFR- β PTK domain) Overexpression	Tel-PDGFR- β : chronic myelomonocytic leukaemia PDGFR- β : glioma
CSF-1R (c-fms)	v-fms from FeSV (p170 ^{gag-fms})	v-fms: Truncated CSF-1R PTK with mutant C-terminal tail Constitutively active c-fms: GOF point mutations Overexpression	v-fms: feline sarcomas c-fms: acute and chronic myelomonocytic leukaemias, monocytic tumours, malignant histiocytosis, endometrial cancer, glioma
Kit/SCFR (c-kit)	v-kit from FeSV (p80 ^{gag-kit})	v-kit: Truncated Kit/SCFR PTK with mutant C-terminal tail Constitutively active c-kit: GOF point mutations and small deletions Overexpression	v-kit: feline fibrosarcomas c-kit: malignant gastrointestinal stromal tumours, acute myeloid leukaemias, myelodysplastic syndromes, mast-cell leukaemia/systemic mastocytosis, seminomas/dysgerminomas, small-cell lung cancer and other carcinomas
Flk2/Flt3		Overexpression Internal tandem gene duplications in JM region	Haematopoietic malignancies
Flt1/VEGFR1		Expression	Tumour angiogenesis
Flk1/VEGFR2		Expression	Tumour angiogenesis
Flt4/VEGFR3		Overexpression	Tumour angiogenesis; vascular tumours (Kaposi's sarcoma, haemangiosarcoma, lymphangiosarcomas)
FGFR1		ZNF198-FGFR1 (t(8; 13) translocation fusing a novel Zn finger protein with the FGFR1 PTK domain) Overexpression Point mutations	ZNF198-FGFR1: acute myelogenous leukaemia (8p11 myeloproliferative syndrome), lymphomas Overexpression: various tumours Point mutations: autosomal skeletal disorders/dysplasias
FGFR2/K-SAM		Overexpression (amplification) and C-terminal truncation	Gastric carcinoma (mammary, prostate carcinomas)
FGFR3		IgH locus/MMSET translocation (t(4; 14) translocation placing FGFR3 PTK downstream of IgH/MMSET). Additional activating FGFR3 point mutations in skeletal dysplasias	Multiple myelomas (achondroplasia, thanatophoric dysplasia and hypochondroplasia)
FGFR4		Overexpression (amplification)	Mammary, ovarian carcinomas
TrkA		Tropomyosin(Tpm)-TrkA (t(1; 1) with N-terminal Tpm sequence fused to TrkA PTK) Tpr-TrkA (t(1; 1) with N-terminal Tpr sequence fused to TrkA PTK) Tfg-TrkA (t(1; 3) with N-terminal Tfg sequence fused to TrkA PTK domain)	Papillary thyroid carcinomas, neuroblastomas
TrkC		Tel-TrkC (t(12; 15) with H-L-H domain of Tel fused to TrkC PTK domain)	Congenital fibrosarcoma, acute myeloid leukaemias
HGFR (c-met)		Tpr-Met (t(1; 7) with N-terminal of Tpr fused to Met PTK domain) Overexpression GOF point mutations	Tpr-Met: Papillary thyroid carcinomas Overexpression: rhabdomyosarcoma, hepatocellular carcinoma GOF point mutations: renal carcinoma
RON		Overexpression/increased kinase activity of splice variants	Colon cancer, hepatocellular cancer
EphA2		Overexpression	Metastasizing malignant melanomas
EphB2		Overexpression	Gastric, oesophageal and colon carcinomas
EphB4		Overexpression	Infiltrating ductal mammary carcinomas
Axl		Overexpression	Acute myeloid leukaemias
TIE/TIE1		Overexpression	Capillary haemangioblastomas, haemangiopericytomas, gastric adenocarcinoma
Tek/TIE2		Expression	Tumour angiogenesis (endothelium)
Ret		Fusions: H4-Ret (PTC1), inversion; R1 α -Ret (PTC2), t(10; 17); ELE1-Ret (PTC3 & PTC4), inversion; RFG5-Ret (PTC5), inversion; HTIF1-Ret (PTC6), t(7; 10); RFG7-Ret (PTC7), t(1; 10); KTN1-Ret (PTC8), t(10; 14); ELKS-Ret, t(10; 12) GOF point mutations: primarily in MEN2A, MEN2B and FMTC (familial), but also in a few sporadic cases	PTCs and ELKS-RET: papillary thyroid carcinomas (5–30% of spontaneous carcinomas, 60–70% of radiation-induced carcinomas (Chernobyl accident)) MEN2A: medullary thyroid carcinoma, parathyroid hyperplasia, pheochromocytoma MEN2B: medullary thyroid carcinoma, pheochromocytoma and enteric mucosal ganglioneuromas. FMTC: medullary thyroid carcinoma

Table 1 Examples of dominant protein-tyrosine kinase oncogenes (continued)

PTK (proto-oncogene)	Viral oncogene* (viral oncoprotein)	Oncogenic alteration	Tumour/cancer types (only the most frequent, mainly human types are described)
ROS (c-ros)	v-ros from avian UR2 SV (p68 ^{gag-ros})	v-ros: Truncated Ros PTK domain. Constitutively active c-ros: Overexpression Rare truncations/point mutations?	v-ros: avian fibrosarcomas c-ros: glioblastomas, astrocytomas
Alk		NPM-Alk (t(2; 5) nucleophosmin fused to Alk PTK domain) Igλ-Alk (t(2; 22) Igλ fused to Alk PTK domain) Other sporadic fusions with tropomyosin, etc.	Non-Hodgkin lymphomas, CD30 ⁺ and CD30 ⁻ anaplastic large-cell lymphoma
Src (c-src)	v-src from RSV (pp60 ^{v-src})	v-src: C-terminal truncation and point mutations (increased kinase activity) c-src: C-terminal truncation (increased kinase activity) Overexpression and/or increased kinase activity	pp60 ^{v-src} : avian sarcomas c-Src truncation: colon cancer c-Src overexpression: mammary and pancreatic cancers, neuroblastomas, others
Fgr (c-fgr)	v-fgr from FeSV (p70 ^{gag-fgr})	v-fgr: C-terminal truncated c-fgr and point mutations (increased kinase activity)	p70 ^{gag-fgr} : feline fibrosarcomas c-Fgr: acute myeloblastic leukaemias, chronic lymphocytic leukaemias, EBV-infected lymphomas; differentiation marker?
Yes (c-yes)	v-yes from ASV (p90 ^{gag-yes})	v-yes: C-terminal truncated c-yes and point mutations (increased kinase activity) Over expression and/or increased kinase activity	p90 ^{gag-yes} : avian sarcomas c-Yes: colon carcinomas, malignant melanomas, other cancers
Lck		TCRβ-Lck (t(1; 7) T-cell receptor-β upstream of Lck causes overexpression) Overexpression GOF point mutations	TCRβ-Lck: T-cell acute lymphocytic leukaemias Lck: chronic lymphocytic leukaemias?
Abl (c-abl)	v-abl from Abelson MLV (p160 ^{gag-abl}) or from PI-FeSV	v-abl: N-terminal (ΔSH3) truncation of Abl Fusions: p190 ^{Bcr-Abl} , (t(9; 22) m-bcr); p210 ^{Bcr-Abl} , (t(9; 22) M-bcr); p230 ^{Bcr-Abl} , (t(9; 22) μ-bcr). M-, m- and μ-bcr: 3 breakpoint cluster regions in BCR. The chimaeric mRNA usually starts with exon a2 of ABL, it never includes exon 1a or 1b. Tel-Abl (t(12; 22) N-terminal (H-L-H region) of Tel fused with Abl exon 2a	p160 ^{gag-abl} : murine acute leukaemias p190 ^{Bcr-Abl} : ~50% of Ph ⁺ acute lymphocytic leukaemias, rarely chronic myelomonocytic leukaemias p210 ^{Bcr-Abl} : chronic myeloid leukaemias, ~30% of Ph ⁺ acute lymphocytic leukaemias p230 ^{Bcr-Abl} : some Ph ⁺ chronic neutrophilic leukaemias Tel-Abl: rare cases of acute lymphocytic leukaemias
Arg		Tel-Arg (t(1; 12))	Acute myeloid leukaemias (rare in M3 and M4Eo subtypes)
Jak1		Overexpression	Various leukaemias
Jak2		Tel-Jak2 (t(9; 12) H-L-H region (N-terminal) of Tel fused with kinase region (C-terminal) of Jak2)	T-cell childhood acute lymphocytic leukaemias, acute myeloid leukaemias, acute lymphocytic leukaemias, atypical chronic myeloid leukaemias
Jak3		Overexpression (increased kinase activity)	Various leukaemias and B-cell malignancies
Fak		Overexpression and/or altered tyrosine kinase activity	Modulation of adhesion, invasion and metastasis of diverse malignancies
Pyk2		Overexpression and/or altered tyrosine kinase activity	Modulation of adhesion, invasion and metastasis of diverse malignancies
Fes (c-fes)	v-fps from ASV, v-fes from FeSV (p130 and p140 ^{gag-fps} , p135 and p140 ^{gag-fes})	The viral gag sequence essential for transforming activity. Viral gag fused to slightly truncated and/or point mutated fps or fes	p130 and p140 ^{gag-fps} : diverse avian sarcomas and myeloid leukaemias p135 and p140 ^{gag-fes} : feline sarcomas c-Fes: no implication in cancers
Brk		Overexpression (increased kinase activity)	Mammary carcinomas
Syk		Downregulation (recessive?)	Potential mammary carcinoma tumour suppressor; PTK-dependent?

*Abbreviations: AVE, avian erythroblastosis virus; FeSV, feline sarcoma virus; UR2 SV, UR2 sarcoma virus; RSV, Rous sarcoma virus; ASV, avian sarcoma virus; Abelson MLV, Abelson murine leukaemia virus; PI-FeSV, Parodi-Irgens feline sarcoma virus. Other abbreviations are defined in the text. Limitations on space prevent the addition of references to the data presented in Table 1.

sequences⁴⁸. The Bcr-Abl-tyrosine kinase domain is activated by formation of homo-oligomeric complexes mediated by the Bcr coiled-coil domain, allowing *trans* autophosphorylation⁵². The transforming effect of Bcr-Abl is mediated by numerous downstream signalling pathways normally activated by RPTKs, including the Ras-Raf-ERK, JAK-STAT and PI(3)K pathways⁵³. The Bcr-Abl-activated pathways are very similar to those activated by Kit/SCFR⁵⁴; the PI(3)K, and perhaps the JAK-STAT, pathways are essential for the mitogenic and anti-apoptotic transforming effects^{55,56}. A requirement for Bcr-Abl kinase activity in transformation is demonstrated by the fact that the small-molecule c-Abl PTK inhibitor STI 571 (ref. 41) not only prevents cell growth of Bcr-Abl-transformed leukaemic cells, but also induces apoptosis in a manner dependent on Stat5-induced upregulation of the Bcl-2-like Bcl-x_L (ref. 57). Cytoplasmic retention of Bcr-Abl is in part kinase-dependent, so treatment of Bcr-Abl-transformed cells with STI 571 results in nuclear import. This observation can be used to enhance cell killing, by taking advantage of the pro-apoptotic effect of nuclear Abl. Treatment of STI 571-inhibited cells with leptomycin B, a nuclear export inhibitor, causes Bcr-Abl accumulation in the nuclear compartment; subsequent removal of STI 571 then results in re-activation of Bcr-Abl and

apoptosis⁴⁸. Although leptomycin B is too neurotoxic for use in treating patients, this rationale might have practical implications for combined use of STI 571 with a less toxic nuclear export inhibitor.

JAKs and STATs

The Janus PTKs (JAKs) have so far been implicated in a limited number of human leukaemias (Table 1), while some of their substrates, Stat3 and Stat5, are found in an activated phosphorylated state in several malignancies. The JAKs are cytoplasmic PTKs (Fig. 2) that mediate signalling primarily downstream of cytokine receptors (which lack catalytic domains), but also of RPTKs (see ref. 58 for a review). The seven known mammalian STATs are latent transcription factors with a central DNA-binding region and a C-terminal SH2 domain. In response to ligand, cytokine receptors become tyrosine-phosphorylated by the constitutively associated JAKs. Some of the phosphorylated residues subsequently bind STATs through their SH2 domains, which become phosphorylated by the oligomerized JAKs on a C-terminal tyrosine residue. This leads to STAT oligomerization through a reciprocal interaction between SH2 and phosphotyrosine⁵⁹. Dimeric STATs are released from the receptors and translocate to the nucleus where they activate transcription. In

addition to activation by JAKs, STATs are also activated (either directly or indirectly) by RPTKs, G-protein-coupled receptors and cytoplasmic PTKs such as Src and Abl⁶⁰.

Ligand-induced STAT tyrosine phosphorylation is a transient and tightly controlled process lasting from minutes to hours⁶¹. Three types of inhibitors of STAT activation are known. The SHP-1 and SHP-2 PTPs directly dephosphorylate JAKs, and an unidentified nuclear PTP has been implicated in tyrosine dephosphorylation of Stat1, which might be important for inactivation and nuclear export of STATs⁶². The cytokine-inducible SH2-containing protein-1 (CIS-1)/suppressor of cytokine signalling-1, -2 and -3 (SOCS-1, -2 and -3)/JAK-binding protein (JBP)/STAT-induced STAT inhibitor-1 (SSI-1) family of proteins is transcriptionally upregulated by STATs, to inhibit STAT signalling. This occurs through SOCS-induced protein degradation, and, for SOCS-1 and -3, through binding of the N-terminal region to the activation loop of JAK kinase⁶³. Finally, the protein inhibitors of activated STATs (PIAS) family of proteins inhibit phosphorylated STAT dimers from DNA binding and hence transcriptional activation. Inactivation of such negative regulators might be involved in tumorigenesis. A small deletion on chromosome 16 containing SOCS-1 has been identified in almost 50% of hepatocellular carcinomas.

Stat3 and Stat5 are overexpressed in some human malignancies. In head and neck cancer, Stat3 is persistently activated as a result of EGFR amplification, and has been shown to be required for v-Src-induced transformation. Moreover, expression of a constitutively active, dimeric Stat3 is transforming. Activation of Stat5 is implicated in human breast cancers⁶⁰. A recurrent translocation involving the N-terminal dimerization domain of the ETS-like transcription factor TEL and Jak2, resulting in dimerization and constitutive activation of Jak2, has been reported in a few cases of T-cell acute lymphocytic leukaemia^{64,65}. In a mouse model, Stat5 is essential for the Tel-Jak2-induced transformation⁶⁶. Undoubtedly, the transforming properties of Stat3 and Stat5 depend on cross-talk with other signal-transduction pathways, in particular the Src and PI(3)K pathways. Thus, Jak1 and Jak2 bind and activate PI(3)K, activated Stat3 has been shown to activate Src and PI(3)K by direct binding through an interaction between phosphotyrosine and SH2, and Stat5 cooperates with PI(3)K in oncogenesis, for example, through induction of Bcl-x_L and inactivation of Bad^{67–69}.

PI(3)K and ribosomal S6 kinase/mTOR signalling

Numerous cytoplasmic protein-serine/threonine kinases, including Raf and ERK in the classical Ras-Raf-MEK-ERK pathway, are involved in cellular proliferation and have been linked to cancer. However, although mutationally activated or overexpressed Raf, Cot/Tpl2 and Mos protein-serine/threonine kinases can transform via the ERK/MAP kinase pathway, no frequently recurring mutations in protein kinases in the MAP kinase pathways have been identified in human malignancies. In contrast, mutations in several proteins in the PI(3)K/Akt and mammalian target of rapamycin (mTOR)/p70^{S6K} signalling pathways, which regulate cell survival, growth and proliferation^{70–74}, are causally involved in a high percentage of common human malignancies, including mammary, prostate and colon carcinomas and malignant brain tumours.

PI(3)K

PI(3)Ks are a family of lipid kinases defined by their ability to phosphorylate the 3'-OH group of the inositol ring in inositol phospholipids. Class I PI(3)Ks are heterodimers made up of a catalytic subunit of relative molecular mass 110,000 (p110) and an adaptor/regulatory subunit. This class is further subdivided into the RPTK-activated subclass IA and the heterotrimeric G-protein-coupled receptor-activated subclass IB. The preferred substrate for class I PI(3)Ks in the intact cell is PtdIns(4,5)P₂. In addition, some of the class I and III PI(3)K members also exhibit protein-serine/threonine kinase activity, the functional role of which is still being explored.

There are currently three isoforms, α , β and δ , of the catalytic p110 subunit, and seven adaptor proteins generated by expression and alternative splicing of three genes, p85 α , p85 β and p55 γ , in the class IA family of PI(3)Ks. Activation of class IA PI(3)Ks can occur by several means. RPTK activation leads to recruitment of PI(3)K, which binds through one or both SH2 domains in the adaptor to specific phosphotyrosine consensus motifs. This leads to allosteric activation of the catalytic subunit. In addition, Ras-GTP can bind directly to an N-terminal region in p110, leading to PI(3)K activation. Activation results in production of PtdIns(3,4,5)P₃ within seconds, and a slightly delayed production of PtdIns(3,4)P₂ through the actions of 5'-inositol lipid phosphatases. The effects of polyphosphoinositides in cells are mediated through the specific binding to at least two lipid-binding protein domains, the FYVE and pleckstrin-homology (PH) domains (see ref. 2 and <http://smart.embl-heidelberg.de/> for protein modules in signalling). FYVE domains bind selectively to PtdIns(3)P, whereas a subgroup of PH domains, containing a highly basic motif, binds PtdIns(3,4)P₂ and PtdIns(3,4,5)P₃. Proteins containing the latter domain are key mediators of class IA PI(3)K signalling. Phosphoinositide-binding PH domains are found in numerous proteins, including the protein-serine/threonine kinases, 3'-phosphoinositide-dependent kinase-1 (PDK-1) and Akt/protein kinase B (PKB), both central for the transforming effects of deregulated PI(3)K activity.

PDK-1 and PKB/Akt

PKB/Akt is the cellular homologue of the transforming viral oncogene v-Akt and bears significant homology to PKA and PKC⁷⁵. The three mammalian isoforms, α , β and γ , all contain an N-terminal PH domain, a central kinase domain with an activation-loop Thr308 phosphorylation site, and a conserved, regulatory serine phosphorylation site, Ser473, near the C terminus. PDK-1 is a Thr308-Akt kinase, and only one mammalian isoform is known⁷⁶. The C-terminal PH domain in PDK-1 binds phospholipids with around tenfold-higher affinity than the Akt PH domain, which probably explains the constitutive localization of PDK-1 at the plasma membrane. The following model for activation of Akt has been established (Fig. 3c): RPTK activation leads to production of PtdIns(3,4,5)P₃ and PtdIns(3,4)P₂ at the inner leaflet of the membrane. Akt interacts with these phospholipids, causing its translocation to the inner membrane, where PDK-1 is located. The interaction of the Akt PH domain with 3'-phosphoinositides is thought to impose conformational changes in Akt, exposing its two main phosphorylation sites. The PH domains might also mediate protein proximity between Akt and PDK-1 through homodimerization. PDK-1, believed to be constitutively active, subsequently phosphorylates Thr308 in Akt, which stabilizes the activation loop in an active conformation. This model is reminiscent of the general model for PTK activation (Fig. 3a–c). Phosphorylation of Thr308 is a prerequisite for kinase activation, but phosphorylation of the C-terminal hydrophobic residue is required as well for full activation of Akt kinase. The Akt Ser473 kinase ('PDK-2') remains to be identified (see ref. 76 for discussion). In a later phase, through unknown mechanisms, activated Akt is translocated to the nucleus where several of its substrates reside⁷⁷.

PDK-1 phosphorylates numerous other AGC kinase members in addition to Akt at the conserved activation-loop Ser or Thr residue, including several PKC isoforms, the serum- and glucocorticoid-induced kinases (SGKs), PKC-related kinase (PRK), p70^{S6K} and p21-activated protein kinase (PAK; see Fig. 4). Results obtained using PDK-1^{-/-} embryonic stem cells indicate that Akt, p70^{S6K} and p90RSK are physiological PDK-1 substrates, while PDK-1 is not required for phosphorylation of PKA, mitogen- and stress-activated protein kinase-1 (MSK-1) and the AMP-activated protein kinase (AMPK; ref. 78). Additional evidence indicates that PKC ζ and PKC δ , but not PRK, are also phosphorylated in the activation loop *in vivo* by PDK-1. The phosphorylation of the activation loop in p70^{S6K} provided a long-sought link between PI(3)K and p70^{S6K} activation.

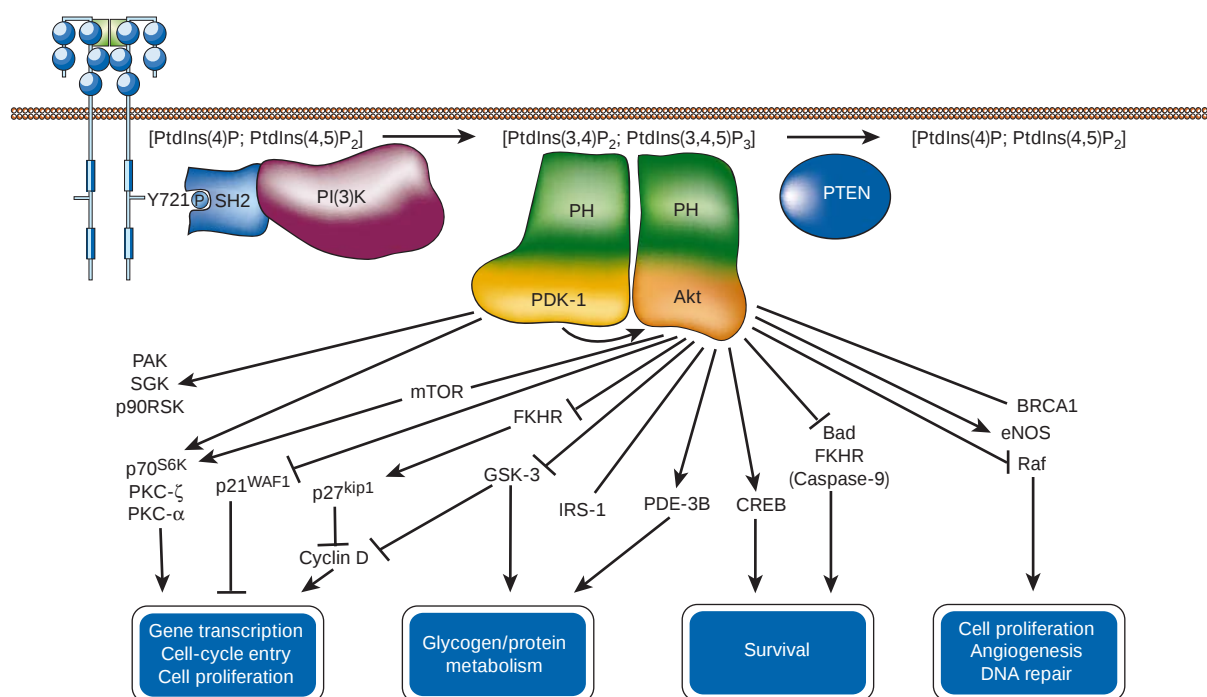


Figure 4 RPTK-induced PI(3)K signalling through PDK-1 and Akt. The figure illustrates signalling from the human Kit/SCFR, but the same general mechanisms apply to most RPTKs. See text for details.

At least 13 Akt substrates have been identified so far in mammalian cells, and they fall into two main classes: regulators of apoptosis on one hand and of cell growth, including protein synthesis and glycogen metabolism, and cell-cycle regulation on the other (Fig. 4). All identified substrates are phosphorylated within the same basic motif, R-X-R-X-X-S/T, which can also be phosphorylated by MAPKAPK-1 and p70^{S6K}. The Akt substrates involved in cell-death regulation include Forkhead transcription factors, the pro-apoptotic Bcl-2 family member Bad, and the cyclic AMP response element-binding protein (CREB). The anti-apoptotic effects of Akt-mediated phosphorylation of these have been extensively reviewed⁷⁹. Glycogen synthase kinase-3 (GSK-3), phosphodiesterase-3B, mTOR, insulin receptor substrate-1 (IRS-1), the Forkhead member FKHR, the cyclin-dependent kinase inhibitor p21^{CIP1/WAF1} and possibly Raf-1 are targets involved in mediating protein synthesis, glycogen metabolism and cell-cycle regulation. GSK-3 is inhibited by Akt phosphorylation. This abolishes phosphorylation of the cytoplasmic signalling molecule β -catenin, causing its stabilization and nuclear translocation. In the nucleus, it associates with T-cell factor/lymphocyte enhancer-binding factor-1 (TCF/LEF-1) to induce the transcription of several genes, including cyclin D1. This results in cell cycle progression through hyperphosphorylation and inactivation of Rb (see article in this issue by Evan and Vousden, pages 342–348). Cyclin D1 is also stabilized in this manner, owing to decreased phosphorylation at a GSK-3 site which promotes proteolytic turnover of cyclin D1⁸⁰. Phosphorylation of p21 by Akt causes its cytoplasmic retention, preventing it from exerting its anti-proliferative effects in the nucleus⁸¹. Phosphorylation of endothelial nitric oxide synthase (eNOS) and breast cancer susceptibility-1 (BRCA1) might regulate angiogenesis and DNA repair, among others (refs 70, 79, and see review in this issue by Hoeijmakers, pages 366–374).

The oncogenic role of deregulated class IA PI(3)Ks and Akt activity is probably accounted for by their ability to induce multiple simultaneous effects on both cell survival and cell cycle/cell growth. Akt β is overexpressed in pancreatic and ovarian carcinomas, and the transforming effect of a constitutively active p110 found in a chicken

tumour virus, p3k, is mediated through Akt⁸². Increased Akt kinase activity is correlated with p53 — an oncogenic mutant of p53 α that induces constitutive PI(3)K activity — and amplification of p110 in ovarian cancer^{83,84}. In addition, numerous human malignancies, including breast cancer, glioblastoma and germ cell tumours, are associated with inactivating mutations in the tumour-suppressor gene PTEN, leading to deregulated hyperactivity of Akt. PTEN is a 3'-phosphoinositide phosphatase, which dephosphorylates the 3'-OH position of the inositol ring in PtdIns(3,4,5)P₃ and PtdIns(3,4)P₂. Consequently, inactivating PTEN mutations lead to increased levels of 3'-phosphoinositides, causing enhanced Akt activity and cellular transformation. Part of the transforming effect of mutant PTEN and, as a consequence, deregulated Akt activity might occur through downregulation of the cyclin-dependent kinase inhibitor p27, a finding highly relevant to human prostate carcinoma⁸⁵. Interestingly, as an apparent exception to these examples, disruption of p110 γ , a class IB PI(3)K, is associated with colorectal cancers⁸⁶. This was reportedly due to upregulation of Bcl-2, cyclin D and CDKs, but not of class IA PI(3)Ks, but the transforming mechanisms remain to be identified.

mTOR and ribosomal S6K

The phosphorylation of the ribosomal p70^{S6K} by PDK-1 and of mammalian TOR (mTOR) by Akt provides mechanistic links between the two pathways, and recently it has become clear that some of the transforming, cell-growth- and cell-cycle-promoting effects of PI(3)K are mediated through the mTOR/p70^{S6K} pathway. mTOR belongs to an evolutionarily conserved family of proteins, including TOR1, TOR2, MEC1, TEL1 and Rad3 in budding yeast, MEI-41 in the fruitfly *Drosophila*, and DNA-dependent protein kinase (DNA-PK), ATM, ATM-related (ATR), transformation/transcription domain-associated protein (TRAPP) and mTOR in mammals. mTOR is also called FK506-binding protein (FKBP)-rapamycin-associated protein (FRAP) in humans and rapamycin and FKBP12 target-1 (RAFT-1) in rats, based on the ability of FKBP-rapamycin complex to bind and inhibit mTOR. Two

mTOR-encoding genes have been identified, but only one has been studied.

Proteins in the mTOR family all have a C-terminal kinase domain with homology to the core kinase domain of PI(3)Ks and PtdIns(4)Ks, but only serine/threonine kinase activity has been demonstrated for these proteins⁷¹. Growth factors stimulate mTOR kinase activity, but the exact regulatory mechanisms are unknown. In response to insulin, Akt phosphorylates two sites in the C terminus of mTOR, but only one of these is a major site *in vivo*. However, this phosphorylation is not required for mTOR to phosphorylate two of its main substrates, eukaryotic initiation factor-4E (eIF-4E)-binding protein (4E-BP) and p70^{S6K}, which are involved in initiation of protein translation. Based on the use of the mTOR inhibitor rapamycin, it is clear that mTOR also regulates transcription of *c-myc* and is involved in activation of Stat3 by phosphorylation of Stat3 at Ser727 and of PKC α and PKC δ by phosphorylation at a conserved residue in their hydrophobic C-terminal motif. These effects might be involved in tumorigenesis^{71,72}. The phosphorylation of 4E-BP causes dissociation from eIF-4E, enabling the latter to participate in cap-dependent initiation, including translation of mRNAs with a highly structured 5'-untranslated region, such as the transcripts encoding *c-Myc* and cyclin D1 involved in cell-cycle progression.

Two ribosomal S6 kinases, S6K1 and S6K2, are known, and they are key regulators of cell growth through control of the protein translational apparatus, in particular ribosomal proteins⁷³. The shorter isoform of S6K1, p70^{S6K}, is largely cytoplasmic and its activation is blocked by rapamycin. p70^{S6K} regulation is complex, and requires hierarchical phosphorylation. Activated p70^{S6K} mediates the effects of mTOR on protein translation through its phosphorylation of the 40S ribosomal protein subunit S6, which drives translation of 5'-terminal oligopyrimidine-rich tract (5'-TOP) mRNAs. These mRNAs encode primarily ribosomal proteins and other protein components of the translational machinery⁷³. However, p70^{S6K} is also involved in cell-cycle regulation. Hence, p70^{S6K} has been linked to PI(3)K-dependent proliferation through upregulation of cyclin D3 and resulting phosphorylation of Rb and p107, which leads to enhanced transcriptional activity of E2F^{87,88}. Transformation by PI(3)K and Akt, but not by numerous other oncoproteins, is dependent on phosphorylation and activation of p70^{S6K} and phosphorylation of 4E-BP-1 by mTOR⁸⁹, thus establishing a clear link between PI(3)K, Akt, p70^{S6K} and mTOR in oncogenesis.

Concluding remarks

Cancer is a multistep process, with accumulation of mutations in tumour-suppressor genes and dominant oncogenes. But the recent development of a series of relatively specific PTK inhibitors, and their ability to inhibit the proliferation of tumour cells expressing the target PTK *in vivo*, shows that inhibition of a deregulated, dominant oncogenic PTK is often enough to slow tumour progression. In consequence, much recent effort has gone into designing and identifying additional PTK inhibitors that are even more potent and specific. To facilitate drug testing there is a clear need for better animal models that precisely reflect the mutations in and pathogenesis of human malignancies. Most of the current refined animal models are based on conditional transgene expression and/or conditional gene knockouts. Not surprisingly, these models often give rise to a completely different tumour spectrum than that found in humans, and at best reflect a very late stage of the oncogenic process where chromosomal rearrangements have resulted in amplifications and deletions. Accordingly, only a few knockout models, in most cases involving deletions of tumour-suppressor genes which provide less attractive drug targets, provide realistic models (see for example, refs 85, 90). Most human cancers are caused initially by somatic point mutations, with only ~1% being due to germline mutations. In order to recapitulate the early events in human oncogenesis, including oncogenic lesions due to mutated PTKs, there is a need to develop mice with inducible/reversible site-directed point mutations.

Note added in proof. Initial results showing the efficacy and safety of STI571 for treating chronic myeloid leukaemia and Kit-positive gastrointestinal stromal tumours have just been published^{91–93}. □

- Hunter, T. Signaling—2000 and beyond. *Cell* **100**, 113–127 (2000).
- Pawson, T. & Nash, P. Protein-protein interactions define specificity in signal transduction. *Genes Dev.* **14**, 1027–1047 (2000).
- Reed, J. C. Dysregulation of apoptosis in cancer. *J. Clin. Oncol.* **17**, 2941–2953 (1999).
- Hunter, T. The Croonian Lecture 1997. The phosphorylation of proteins on tyrosine: its role in cell growth and disease. *Phil. Trans. R. Soc. Lond. B* **353**, 583–605 (1998).
- Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
- Plowman, G. D., Sudarsanam, S., Bingham, J., Whyte, D. & Hunter, T. The protein kinases of *Caenorhabditis elegans*: a model for signal transduction in multicellular organisms. *Proc. Natl Acad. Sci. USA* **96**, 13603–13610 (1999).
- Robinson, D. R., Wu, Y. M. & Lin, S. F. The protein tyrosine kinase family of the human genome. *Oncogene* **19**, 5548–5557 (2000).
- Futreal, P. A. *et al.* Cancer and genomics. *Nature* **409**, 850–852 (2001).
- Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **103**, 211–225 (2000).
- Blume-Jensen, P. & Hunter, T. In *Encyclopedia of Cancer* (ed. Bertino, J. R.) (Academic, San Diego, in the press).
- Heldin, C.-H. Dimerization of cell surface receptors in signal transduction. *Cell* **80**, 213–223 (1995).
- Hubbard, S. R. & Till, J. H. Protein tyrosine kinase structure and function. *Annu. Rev. Biochem.* **69**, 373–398 (2000).
- Hubbard, S. R., Mohammadi, M. & Schlessinger, J. Autoregulatory mechanisms in protein-tyrosine kinases. *J. Biol. Chem.* **273**, 11987–11990 (1998).
- Jiang, G. & Hunter, T. Receptor signaling: when dimerization is not enough. *Curr. Biol.* **9**, R568–R571 (1999).
- Shewchuk, L. M. *et al.* Structure of the Tie2 RTK domain. Self-inhibition by the nucleotide binding loop, activation loop, and C-terminal tail. *Structure Fold. Des.* **8**, 1105–1113 (2000).
- Binns, K. L., Taylor, P. P., Sicheri, F., Pawson, T. & Holland, S. J. Phosphorylation of tyrosine residues in the kinase domain and juxtamembrane region regulates the biological and catalytic activities of Eph receptors. *Mol. Cell. Biol.* **20**, 4791–4805 (2000).
- Irusta, P. M. & DiMaio, D. A single amino acid substitution in a WW-like domain of diverse members of the PDGF receptor subfamily of tyrosine kinases causes constitutive receptor activation. *EMBO J.* **17**, 6912–6923 (1998).
- Hirota, S. *et al.* Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. *Science* **279**, 577–580 (1998).
- Meshinchi, S. *et al.* Prevalence and prognostic significance of Flt3 internal tandem duplication in pediatric acute myeloid leukemia. *Blood* **97**, 89–94 (2001).
- Harari, D. & Yarden, Y. Molecular mechanisms underlying ErbB2/HER2 action in breast cancer. *Oncogene* **19**, 6102–6114 (2000).
- Jiang, S. M. The RET proto-oncogene in human cancers. *Oncogene* **19**, 5590–5597 (2000).
- Santoro, M. *et al.* Activation of RET as a dominant transforming gene by germline mutations of MEN2A and MEN2B. *Science* **267**, 381–383 (1995).
- Zhou, S. *et al.* Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. *Nature* **373**, 536–539 (1995).
- Boccardi, R. *et al.* The multiple endocrine neoplasia type 2B point mutation switches the specificity of the Ret tyrosine kinase towards cellular substrates that are susceptible to interact with Crk and Nck. *Oncogene* **15**, 2257–2265 (1997).
- Murakami, H. *et al.* Enhanced phosphatidylinositol 3-kinase activity and high phosphorylation state of its downstream signalling molecules mediated by ret with the MEN 2B mutation. *Biochem. Biophys. Res. Commun.* **262**, 68–75 (1999).
- Smith-Hicks, C. L., Sizer, K. C., Powers, J. F., Tischler, A. S. & Costantini, F. C-cell hyperplasia, pheochromocytoma and sympathoadrenal malformation in a mouse model of multiple endocrine neoplasia type 2B. *EMBO J.* **19**, 612–622 (2000).
- van Weering, D. H. & Bos, J. L. Signal transduction by the receptor tyrosine kinase Ret. *Recent Results Cancer Res.* **154**, 271–281 (1998).
- Ashman, L. K. The biology of stem cell factor and its receptor C-kit. *Int. J. Biochem. Cell Biol.* **31**, 1037–1051 (1999).
- Besmer, P. The kit ligand encoded at the murine Steel locus: a pleiotropic growth and differentiation factor. *Curr. Opin. Cell Biol.* **3**, 939–946 (1991).
- Blume-Jensen, P. & Hunter, T. In *Cancer Research—An Encyclopedic Reference* (ed. Schwab, M.) (Springer, in the press).
- Kozlowski, M. *et al.* SHP-1 binds and negatively modulates the c-Kit receptor by interaction with tyrosine 569 in the c-Kit juxtamembrane domain. *Mol. Cell. Biol.* **18**, 2089–2099 (1998).
- Lennartsson, J. *et al.* Phosphorylation of Shc by Src family kinases is necessary for stem cell factor receptor/c-kit mediated activation of the Ras/MAP kinase pathway and c-fos induction. *Oncogene* **18**, 5546–5553 (1999).
- Taniguchi, M. *et al.* Effect of c-kit mutation on prognosis of gastrointestinal stromal tumors. *Cancer Res.* **59**, 4297–4300 (1999).
- Schmidt, L. *et al.* Germline and somatic mutations in the tyrosine kinase domain of the MET proto-oncogene in papillary renal carcinomas. *Nature Genet.* **16**, 68–73 (1997).
- Chian, R. *et al.* PI3 kinase mediates transformation of hematopoietic cells by the V816 c-kit mutant. *Exp. Hematol.* **28**, 1491 (2000).
- Blume-Jensen, P., Janknecht, R. & Hunter, T. The kit receptor promotes cell survival via activation of PI 3-kinase and subsequent Akt-mediated phosphorylation of Bad on Ser136. *Curr. Biol.* **8**, 779–782 (1998).
- Chen, H. E., Chang, S., Trub, T. & Neel, B. G. Regulation of colony-stimulating factor 1 receptor signaling by the SH2 domain-containing tyrosine phosphatase SHPTP1. *Mol. Cell. Biol.* **16**, 3685–3697 (1996).
- Paulson, R. F., Vesely, S., Siminovich, K. A. & Bernstein, A. Signalling by the W/Kit receptor tyrosine kinase is negatively regulated *in vivo* by the protein tyrosine phosphatase Shp1. *Nature Genet.* **13**, 309–315 (1996).
- Piao, X., Paulson, R., van der Geer, P., Pawson, T. & Bernstein, A. Oncogenic mutation in the Kit receptor tyrosine kinase alters substrate specificity and induces degradation of the protein tyrosine phosphatase SHP-1. *Proc. Natl Acad. Sci. USA* **93**, 14665–14669 (1996).

40. Beghini, A. *et al.* RNA hyperediting and alternative splicing of hematopoietic cell phosphatase (PTPN6) gene in acute myeloid leukemia. *Hum. Mol. Genet.* **9**, 2297–2304 (2000).
41. Druker, B. J. *et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nature Med.* **2**, 561–566 (1996).
42. Heinrich, M. C. *et al.* Inhibition of c-kit receptor tyrosine kinase activity by STI 571, a selective tyrosine kinase inhibitor. *Blood* **96**, 925–932 (2000).
43. Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170–173 (1976).
44. Bjorge, J. D., Jakymiw, A. & Fujita, D. J. Selected glimpses into the activation and function of src kinase. *Oncogene* **19**, 5620–5635 (2000).
45. Sicheri, F. & Kuriyan, J. Structures of Src-family tyrosine kinases. *Curr. Opin. Struct. Biol.* **7**, 777–785 (1997).
46. Irby, R. B. *et al.* Activating SRC mutation in a subset of advanced human colon cancers. *Nature Genet.* **21**, 187–190 (1999).
47. Nowell, P. & Hungerford, D. A minute chromosome in human chronic granulocytic leukemia. *Science* **132**, 1497 (1960).
48. Vigneri, P. & Wang, J. Y. Induction of apoptosis in chronic myelogenous leukemia cells through nuclear entrapment of BCR-ABL tyrosine kinase. *Nature Med.* **7**, 228–234 (2001).
49. Wang, J. Y. Regulation of cell death by the abl tyrosine kinase. *Oncogene* **19**, 5643–5650 (2000).
50. Stiewe, T. & Putzer, B. M. Role of the p53-homologue p73 in E2F1-induced apoptosis. *Nature Genet.* **26**, 464–469 (2000).
51. Plattner, R., Kadlec, L., DeMali, K. A., Kazlauskas, A. & Pendergast, A. M. c-Abl is activated by growth factors and Src family kinases and has a role in the cellular response to PDGF. *Genes Dev.* **13**, 2400–2411 (1999).
52. McWhirter, J. R., Galasso, D. L. & Wang, J. Y. A coiled-coil oligomerization domain of Bcr is essential for the transforming function of Bcr-Abl oncoproteins. *Mol. Cell. Biol.* **13**, 7587–7595 (1993).
53. Deininger, M. W., Goldman, J. M. & Melo, J. V. The molecular biology of chronic myeloid leukemia. *Blood* **96**, 3343–3356 (2000).
54. Hallek, M. *et al.* Interaction of the receptor tyrosine kinase p145c-kit with the p210bcr/abl kinase in myeloid cells. *Br. J. Haematol.* **94**, 5–16 (1996).
55. Gesbert, F., Sellers, W. R., Signoretti, S., Loda, M. & Griffin, J. D. BCR/ABL regulates expression of the cyclin-dependent kinase inhibitor p27Kip1 through the phosphatidylinositol 3-Kinase/AKT pathway. *J. Biol. Chem.* **275**, 39223–39230 (2000).
56. Skorski, T. *et al.* Transformation of hematopoietic cells by BCR/ABL requires activation of a PI-3K/Akt-dependent pathway. *EMBO J.* **16**, 6151–6161 (1997).
57. Horita, M. *et al.* Blockade of the Bcr-Abl kinase activity induces apoptosis of chronic myelogenous leukemia cells by suppressing signal transducer and activator of transcription 5-dependent expression of Bcl-xL. *J. Exp. Med.* **191**, 977–984 (2000).
58. Ihle, J. N., Nosaka, T., Thierfelder, W., Quelle, F. W. & Shimoda, K. Jaks and Stats in cytokine signaling. *Stem Cells* **15**(Suppl. 1), 105–111; discussion, 112 (1997).
59. Darnell, J. E. STATs and gene regulation. *Science* **277**, 1630–1635 (1997).
60. Bromberg, J. F. Activation of STAT proteins and growth control. *BioEssays* **23**, 161–169 (2001).
61. Bowman, T., Garcia, R., Turkson, J. & Jove, R. STATs in oncogenesis. *Oncogene* **19**, 2474–2488 (2000).
62. Haspel, R. L. & Darnell, J. E. A nuclear protein tyrosine phosphatase is required for the inactivation of Stat1. *Proc. Natl Acad. Sci. USA* **96**, 10188–10193 (1999).
63. Yasukawa, H., Sasaki, A. & Yoshimura, A. Negative regulation of cytokine signaling pathways. *Annu. Rev. Immunol.* **18**, 143–164 (2000).
64. Peeters, P. *et al.* Fusion of TEL, the ETS-variant gene 6 (ETV6), to the receptor-associated kinase JAK2 as a result of t(9;12) in a lymphoid and t(9;15;12) in a myeloid leukemia. *Blood* **90**, 2535–2540 (1997).
65. Lacronique, V. *et al.* A TEL-JAK2 fusion protein with constitutive kinase activity in human leukemia. *Science* **278**, 1309–1312 (1997).
66. Schwaller, J. *et al.* Stat5 is essential for the myelo- and lymphoproliferative disease induced by TEL/JAK2. *Mol. Cell* **6**, 693–704 (2000).
67. Rosa Santos, S. C., Dumon, S., Mayeux, P., Gisselbrecht, S. & Gouilleux, F. Cooperation between STAT5 and phosphatidylinositol 3-kinase in the IL-3-dependent survival of a bone marrow derived cell line. *Oncogene* **19**, 1164–1172 (2000).
68. Migone, T. S. *et al.* Functional cooperation of the interleukin-2 receptor beta chain and Jak1 in phosphatidylinositol 3-kinase recruitment and phosphorylation. *Mol. Cell. Biol.* **18**, 6416–6422 (1998).
69. Pfeffer, L. M. *et al.* STAT3 as an adapter to couple phosphatidylinositol 3-kinase to the IFNAR1 chain of the type I interferon receptor. *Science* **276**, 1418–1420 (1997).
70. Vanhaesebroeck, B. & Alessi, D. R. The PI3K-PDK1 connection: more than just a road to PKB. *Biochem. J.* **346**(3), 561–576 (2000).
71. Schmelzle, T. & Hall, M. N. TOR, a central controller of cell growth. *Cell* **103**, 253–262 (2000).
72. Pyronnet, S. & Sonenberg, N. Cell-cycle-dependent translational control. *Curr. Opin. Gen. Dev.* **11**, 13–18 (2001).
73. Dufner, A. & Thomas, G. Ribosomal S6 kinase signaling and the control of translation. *Exp. Cell Res.* **253**, 100–109 (1999).
74. Blume-Jensen, P. & Hunter, T. in *Encyclopedia of Cancer* (ed. Bertino, J. R.) (Academic, San Diego, in the press).
75. Bellacosa, A. *et al.* Structure, expression and chromosomal mapping of c-akt: relationship to v-akt and its implications. *Oncogene* **8**, 745–754 (1993).
76. Tokar, A. & Newton, A. C. Cellular signaling: pivoting around PDK-1. *Cell* **103**, 185–188 (2000).
77. Meier, R. & Hemmings, B. A. Regulation of protein kinase B. *J. Recept. Signal. Transduct. Res.* **19**, 121–128 (1999).
78. Williams, M. R. *et al.* The role of 3-phosphoinositide-dependent protein kinase 1 in activating AGC kinases defined in embryonic stem cells. *Curr. Biol.* **10**, 439–448 (2000).
79. Datta, S. R., Brunet, A. & Greenberg, M. E. Cellular survival: a play in three Acts. *Genes Dev.* **13**, 2905–2927 (1999).
80. Diehl, J. A., Cheng, M., Roussel, M. F. & Sherr, C. J. Glycogen synthase kinase-3 β regulates cyclin D1 proteolysis and subcellular localization. *Genes Dev.* **12**, 3499–3511 (1998).
81. Zhou, B. P. *et al.* Cytoplasmic localization of p21Cip1/WAF1 by Akt-induced phosphorylation in HER-2/neu-overexpressing cells. *Nature Cell Biol.* **3**, 245–252 (2001).
82. Chang, H. W. *et al.* Transformation of chicken cells by the gene encoding the catalytic subunit of PI 3-kinase. *Science* **276**, 1848–1850 (1997).
83. Shayesteh, L. *et al.* PIK3CA is implicated as an oncogene in ovarian cancer. *Nature Genet.* **21**, 99–102 (1999).
84. Jimenez, C. *et al.* Identification and characterization of a new oncogene derived from the regulatory subunit of phosphoinositide 3-kinase. *EMBO J.* **17**, 743–753 (1998).
85. Di Cristofano, A., De Acetis, M., Koff, A., Cordon-Cardo, C. & Pandolfi, P. P. Pten and p27^{Kip1} cooperate in prostate cancer tumor suppression in the mouse. *Nature Genet.* **27**, 222–224 (2001).
86. Sasaki, T. *et al.* Colorectal carcinomas in mice lacking the catalytic subunit of PI(3)K. *Nature* **406**, 897–902 (2000).
87. Brennan, P., Babbage, J. W., Thomas, G. & Cantrell, D. p70(s6k) integrates phosphatidylinositol 3-kinase and rapamycin-regulated signals for E2F regulation in T lymphocytes. *Mol. Cell. Biol.* **19**, 4729–4738 (1999).
88. Feng, L. X., Ravindranath, N. & Dym, M. Stem cell factor/c-kit up-regulates cyclin D3 and promotes cell cycle progression via the phosphoinositide 3-kinase/p70 S6 kinase pathway in spermatogonia. *J. Biol. Chem.* **275**, 25572–25576 (2000).
89. Masahiro, A., Blazek, E. & Vogt, P. K. A role of the kinase mTOR in cellular transformation induced by the oncoproteins P3k and Akt. *Proc. Natl Acad. Sci. USA* **98**, 136–141 (2001).
90. Cichowski, K. & Jacks, T. NF1 tumor suppressor gene function: narrowing the GAP. *Cell* **104**, 593–604 (2001).
91. Druker, B. J. *et al.* Efficacy and safety of a specific inhibitor of the BCR-Abl tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344**, 1031–1037 (2001).
92. Druker, B. J. *et al.* Activity of a specific inhibitor of the BCR-Abl tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *N. Engl. J. Med.* **344**, 1038–1042 (2001).
93. Joensuu, H. *et al.* Effect of the tyrosine kinase inhibitor STI571 in a patient with a metastatic gastrointestinal stromal tumor. *N. Engl. J. Med.* **344**, 1052–1056 (2001).

Acknowledgements

We apologize to all colleagues, including those instrumental in generating the large amount of data compiled in Table 1, whose papers were not cited owing to space limitations. We thank M. R. Stratton and G. Pao for providing inspiration for Table 1, and J. Levenson for critical comments and suggestions. P.B.-J. is a Special Fellow of the Leukemia and Lymphoma Society of America and T.H. is a Frank and Else Schilling American Cancer Society Professor.

Genome maintenance mechanisms for preventing cancer

Jan H. J. Hoeijmakers

MGC Department of Cell Biology and Genetics, Centre for Biomedical Genetics, Erasmus University, PO Box 1738, 3000DR Rotterdam, The Netherlands (e-mail: Hoeijmakers@gen.fgg.eur.nl)

The early notion that cancer is caused by mutations in genes critical for the control of cell growth implied that genome stability is important for preventing oncogenesis. During the past decade, knowledge about the mechanisms by which genes erode and the molecular machinery designed to counteract this time-dependent genetic degeneration has increased markedly. At the same time, it has become apparent that inherited or acquired deficiencies in genome maintenance systems contribute significantly to the onset of cancer. This review summarizes the main DNA caretaking systems and their impact on genome stability and carcinogenesis.

Cancer is a disease of our genes. Over time, DNA accumulates changes that activate proto-oncogenes and inactivate tumour-suppressor genes. The genetic instability driving tumorigenesis is fuelled by DNA damage and by errors made by the DNA machinery. However, 'spontaneous' mutations are insufficient to explain the lifetime cancer risk¹. Indeed, numerous links have been identified between oncogenesis and acquired or inherited faulty genome guardians that cause a 'mutator' phenotype, highlighting the key role of DNA protection systems in tumour prevention. Here I focus on the main DNA maintenance mechanisms operating in mammals — nucleotide- and base-excision repair, homologous recombination, end joining, mismatch repair and telomere metabolism — and their relevance for cancer.

A plethora of damages in DNA

The physicochemical constitution of our genes does not guarantee life-long stability or proper function. A perplexing diversity of lesions arises in DNA from three main causes. First, environmental agents such as the ultraviolet (UV) component of sunlight, ionizing radiation and numerous genotoxic chemicals cause alterations in DNA structure, which, if left unrepaired, may lead to mutations that enhance cancer risk. A pronounced example is exposure to genotoxic compounds in cigarette smoke, which are responsible for the most frequent cancer in Western men. Second, (by)products of normal cellular metabolism constitute a permanent enemy to DNA integrity from within. These include reactive oxygen species (superoxide anions, hydroxyl radicals and hydrogen peroxide) derived from oxidative respiration and products of lipid peroxidation. Over 100 oxidative modifications have been identified in DNA². Evolution has invested significantly in reducing the price of its own metabolism by implementing an intricate antioxidant defence system composed of enzymatic (superoxide dismutase, catalase, glutathione peroxidase and peroxyredoxins) and low-molecular-mass scavengers (such as glutathione)³. Finally, some chemical bonds in DNA tend to spontaneously disintegrate under physiological conditions. Hydrolysis of nucleotide residues leaves non-instructive abasic sites. Spontaneous or induced deamination of cytosine, adenine, guanine or 5-methylcytosine converts these bases to the

miscoding uracil, hypoxanthine, xanthine and thymine, respectively⁴. Figure 1a summarizes some of the most common types of DNA damage and their sources.

The consequences of DNA injury

The outcome of DNA damage is diverse and generally adverse (Fig. 1b). Acute effects arise from disturbed DNA metabolism, triggering cell-cycle arrest or cell death. Long-term effects result from irreversible mutations contributing to oncogenesis.

Many lesions block transcription, which in effect inactivates every gene containing damage on the transcribed strand — an outcome directly related to gene length. This has elicited the development of a dedicated repair system, transcription-coupled repair (TCR), which displaces or removes the stalled RNA polymerase and assures high-priority repair. Transcriptional stress, arising from persistent blockage of RNA synthesis, constitutes an efficient trigger for p53-dependent apoptosis (see ref. 5 and the article in this issue by Evan and Vousden, pages 342–348), which may be a significant anti-cancer mechanism.

Lesions also interfere with DNA replication. Recently, a growing class of DNA polymerases, numbered ζ to κ , was discovered which seems devoted specifically to overcoming damage-induced replicational stress^{6,7}. These special polymerases take over temporarily from the blocked replicative DNA polymerase- δ/ϵ (pol δ/ϵ), and possibly from pol α (Fig. 2, follow upper strand). They have more flexible base-pairing properties permitting translesion synthesis, with each polymerase probably designed for a specific category of injury. The number of polymerases preferring damaged templates currently exceeds that for undamaged DNA, which illustrates the magnitude of the problem. But this solution generally comes at the expense of a higher error rate. In fact, this process is responsible for most of damage-induced point mutations⁸ and is thus particularly relevant for oncogenesis. Nevertheless, translesion polymerases still protect the genome. For instance, inherited defects in pol- η , which specializes in relatively error-free bypassing of UV-induced cyclobutane pyrimidine dimers, cause the variant form of the skin cancer-prone disorder xeroderma pigmentosum^{9,10}. In the yeast *Saccharomyces cerevisiae*, a second, probably even more important pathway exists that allows error-free bypass of lesions⁸. This mechanism is based on reinitiation of DNA

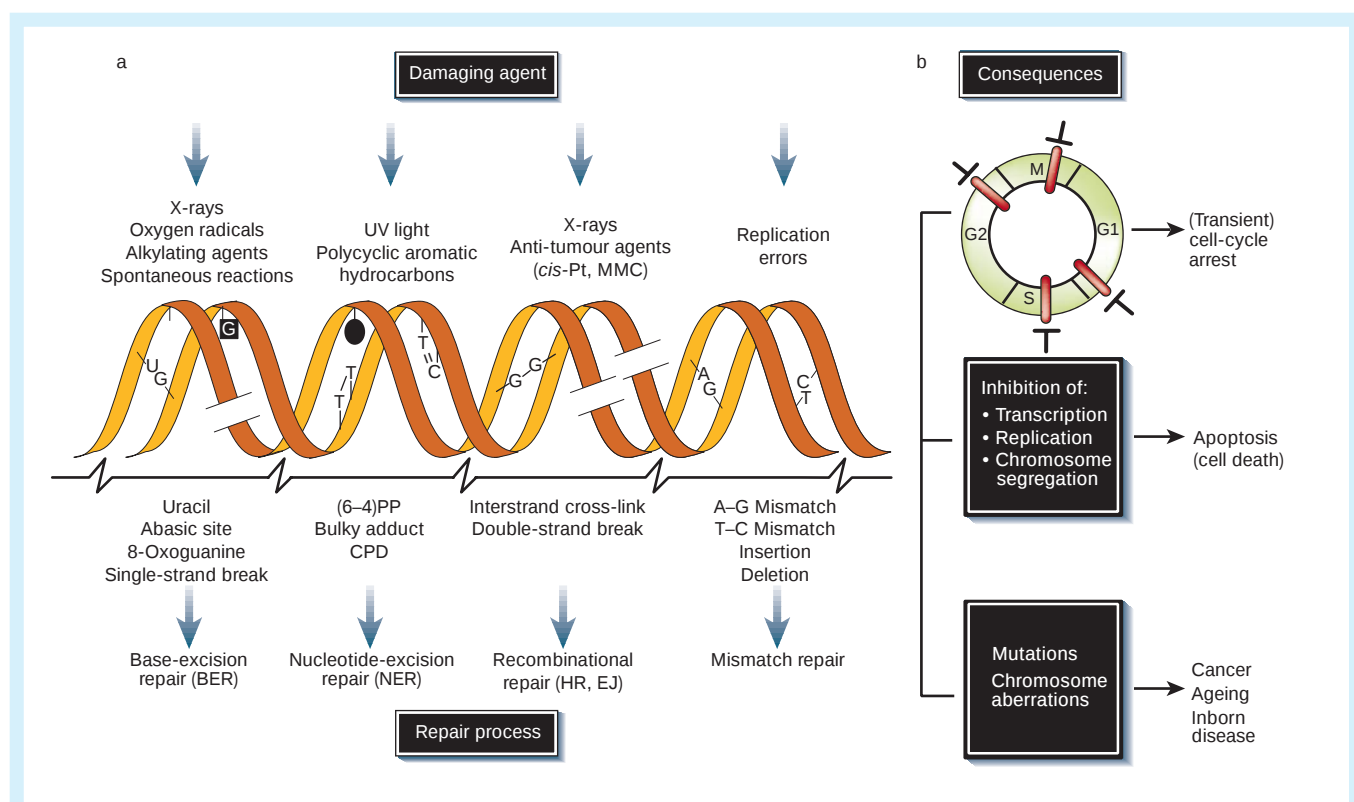


Figure 1 DNA damage, repair mechanisms and consequences. **a**, Common DNA damaging agents (top); examples of DNA lesions induced by these agents (middle); and most relevant DNA repair mechanism responsible for the removal of the lesions (bottom). **b**, Acute effects of DNA damage on cell-cycle progression, leading to transient arrest in the G1, S, G2 and M phases (top), and on DNA metabolism (middle). Long-term consequences of DNA injury (bottom) include permanent changes in the DNA sequence (point mutations affecting single genes or chromosome aberrations which may involve multiple genes) and their biological effects. Abbreviations: *cis*-Pt and MMC, cisplatin and mitomycin C, respectively (both DNA-crosslinking agents); (6-4)PP and CPD, 6-4 photoproduct and cyclobutane pyrimidine dimer, respectively (both induced by UV light); BER and NER, base- and nucleotide-excision repair, respectively; HR, homologous recombination; EJ, end joining.

replication downstream of the blocking injury. The resulting gap is filled in by recombinational replication, using the newly synthesized complementary strand as a template and ignoring the original lesion-containing one (Fig. 2, follow lower strand). Yeast proteins implicated in this process, such as the Ubc13/Mms2 complex, are conserved all the way to mammals. Thus, this largely unexplored system undoubtedly exists in humans and may be important in carcinogenesis. The endpoint of both of these pathways is that damage persists and — when unrepaired — will cause similar problems in subsequent rounds of replication. This is particularly relevant for damage that is not efficiently recognized by any mammalian repair process, such as cyclobutane pyrimidine dimers.

Double-strand DNA breaks (DSBs) induced by X-rays, chemicals or during replication of single-strand breaks (SSBs) and presumably during repair of interstrand crosslinks are particularly relevant for the recombination machinery. Cells with specialized DNA recombination activities, such as B- and T-cells, may be very sensitive to DSBs when they are rearranging their immunoglobulin or T-cell-receptor genes. This explains the frequent involvement of these genetic loci in oncogenic translocations in leukaemia and lymphomas and the preferential induction of these cancers by ionizing irradiation. DSBs also pose problems during mitosis, as intact chromosomes are a prerequisite for proper chromosome segregation during cell division. Thus, these lesions frequently induce various sorts of chromosomal aberrations, including aneuploidy, deletions (loss of heterozygosity) and chromosomal translocations — events which are all intimately associated with carcinogenesis.

The cell-cycle machinery somehow senses genome injury and arrests at specific checkpoints in G1, S, G2 and M to allow repair of lesions before they are converted into permanent mutations

(reviewed in ref. 11). Lesion detection may occur by blocked transcription, replication or specialized sensors. When damage is too significant, a cell may opt for the ultimate mode of rescue by initiating apoptosis at the expense of a whole cell (see review by Evan and Vousden, pages 342–348).

DNA damage repair systems

In view of the plethora of types of lesions, no single repair process can cope with all kinds of damage. Instead, evolution has moulded a tapestry of sophisticated, interwoven DNA repair systems that as a whole cover most (but not all) of the insults inflicted on a cell's vital genetic information. Inherited defects in any of these pathways in general predisposes to malignancy (Table 1). Because the problem of DNA damage has existed *ab initio*, DNA repair systems must have arisen early in evolution. This explains why all known repair pathways are highly conserved (usually across the pro/eukaryotic evolutionary border). At least four main, partly overlapping damage repair pathways operate in mammals — nucleotide-excision repair (NER), base-excision repair (BER), homologous recombination and end joining^{12,13}. The division of tasks between them can be roughly defined as follows (see also Fig. 1a).

NER deals with the wide class of helix-distorting lesions that interfere with base pairing and generally obstruct transcription and normal replication. Small chemical alterations of bases are targeted by BER. These lesions may or may not impede transcription and replication, although they frequently miscode. BER is therefore particularly relevant for preventing mutagenesis. Most NER lesions arise from exogenous sources (except for some oxidative lesions), whereas BER is mostly, but not exclusively, concerned with damage of endogenous origin. Lesions for these two repair processes affect only

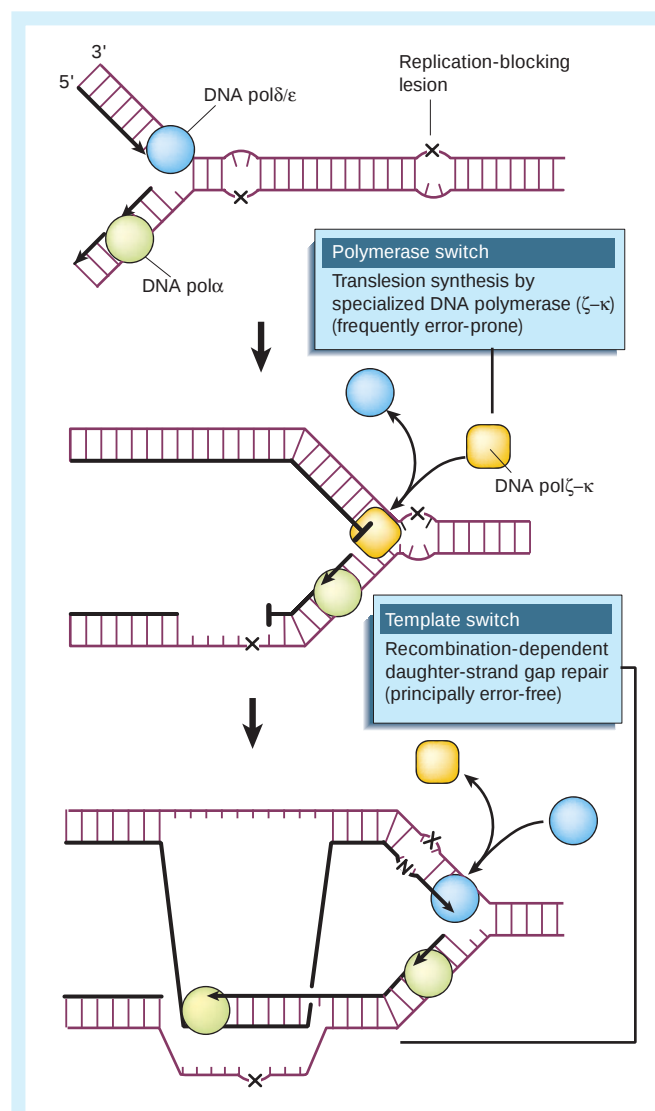


Figure 2 Mechanisms of replicational bypass of DNA lesions. Lesions in the DNA template (indicated by an 'X') may be bypassed by the replication apparatus in two different ways: DNA polymerase switch (upper strand) and template switch (lower strand). In the DNA polymerase switch, the regular DNA polymerase (in this case $\text{pol}\delta/\epsilon$, carrying out leading-strand synthesis) is arrested at the site of the damage. A specific translesion polymerase ($\text{pol}\zeta-\kappa$), or a combination of these polymerases, takes over synthesis to bypass the injured site, after which the regular polymerase continues. This process can be highly error-prone. In the template switch (model), the regular DNA polymerase (in this case $\text{pol}\alpha$, responsible for lagging-strand synthesis) is arrested at a damaged site. The resulting gap in the newly synthesized strand is filled in using the undamaged, newly synthesized leading strand via recombinational strand exchange (or alternatively by fork regression and annealing of the new strand, not shown). This mechanism may involve specific factors as well as members of the RAD52 family implicated in homologous recombination repair. In principle, this mode of lesion bypass is error-free. Note that in both of these processes the lesion remains and that the two scenarios may apply to both strands.

one of the DNA strands. In a 'cut-and-patch'-type reaction, the injury (with or without some flanking sequences) is taken out and the resulting single-stranded gap is filled in using the intact complementary strand as template.

DSBs are more problematic, as both strands are affected. To properly heal such breaks the cell has to know which ends belong together, a difficult task given the size of the mammalian genome. Two pathways, homologous recombination and end joining (and presumably additional back-up systems), were developed for solving

Table 1 Human syndromes with defective genome maintenance

Syndrome	Affected maintenance mechanism	Main type of genome instability	Major cancer predisposition
Xeroderma pigmentosum	NER ($\pm$ TCR)	Point mutations	UV-induced skin cancer
Cockayne syndrome	TCR	Point mutations	None*
Trichothiodystrophy	NER / TCR	Point mutations	None*
Ataxia telangiectasia (AT)	DSB response/repair	Chromosome aberrations	Lymphomas
AT-like disorder	DSB response/repair	Chromosome aberrations	Lymphomas
Nijmegen breakage syndrome	DSB response/repair	Chromosome aberrations	Lymphomas
BRCA 1/BRCA2	HR	Chromosome aberrations	Breast (ovarian) cancer
Werner syndrome	HR?/TLS?	Chromosome aberrations	Various cancers
Bloom syndrome	HR?	Chromosome aberrations (SCE $\uparrow$)	Leukaemia, lymphoma, others
Rothmund–Thomson syndrome	HR?	Chromosome aberrations	Osteosarcoma
Ligase IV deficiency $\dagger$	EJ	Recombination fidelity	Leukaemia(?)
HNPCC	MMR	Point mutations	Colorectal cancer
Xeroderma pigmentosum variant	TLS $\ddagger$	Point mutations	UV-induced skin cancer

*Defect in transcription-coupled repair triggers apoptosis, which may protect against UV-induced cancer.

$\dagger$ One patient with leukaemia and radiosensitivity described with active-site mutation in ligase IV.

$\ddagger$ Specific defect in relatively error-free bypass replication of UV-induced cyclobutane pyrimidine dimers.

Abbreviations: BER, base-excision repair; DSB, double-strand break; HNPCC, hereditary non-polyposis colorectal cancer; HR, homologous recombination; MMR, mismatch repair; NER, nucleotide-excision repair; SCE, sister-chromatid exchange; TCR, transcription-coupled repair; TLS, translesion synthesis.

the DSB problem. Homologous recombination seems to dominate in S and G2 when the DNA is replicated, providing a pristine second copy of the sequence (sister chromatid) for aligning the breaks. In contrast, the less-accurate end joining is most relevant in the G1 phase of the cell cycle, when a second copy is not available¹⁴.

Finally, some single repair proteins directly revert certain injuries, such as O⁶-methylguanine methyltransferase, which removes O⁶-methyl guanine. This highly mutagenic lesion permits base pairing with both C or T and is capable of fooling the mismatch repair system into triggering futile rounds of mismatch removal and subsequent reincorporation of the erroneous base by repair replication. The dedicated methyl transferase specifically removes the non-native methyl group from the guanine residue and transfers it to an internal cysteine. However, in doing so, the protein irreversibly inactivates itself³. This illustrates how in some situations an entire protein may be sacrificed for the repair of a single damaged base. Below I describe the four main multi-step damage repair processes in mammals and their relevance for preventing cancer.

Nucleotide-excision repair and transcription-coupled repair

Of all repair systems, NER is the most versatile in terms of lesion recognition. Two NER subpathways exist with partly distinct substrate specificity: global genome NER (GG-NER) surveys the entire genome for distorting injury, and transcription-coupled repair (TCR) focuses on damage that blocks elongating RNA polymerases¹⁵. Box 1 presents the most likely mechanisms of action for these pathways (and see refs 16, 17).

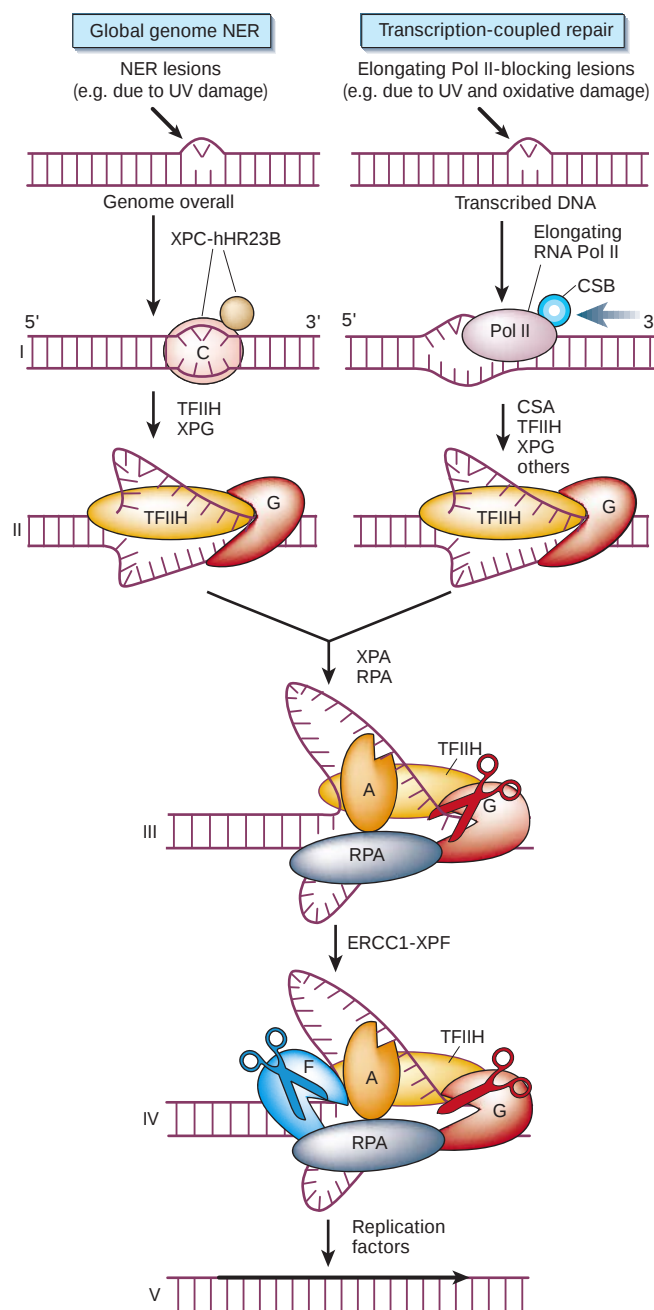
NER, TCR and cancer

At least three syndromes are associated with inborn defects in NER (Table 1): xeroderma pigmentosum, Cockayne syndrome and trichothiodystrophy (TTD), all characterized by exquisite sun sensitivity^{18,19}. The prototype repair disorder, xeroderma pigmentosum,

Box 1

Model for mechanism of global genome nucleotide-excision repair and transcription-coupled repair

The GG-NER-specific complex XPC-hHR23B screens first on the basis of disrupted base pairing⁵³, instead of lesions per se. This explains why mildly distorting injury such as cyclobutane pyrimidine dimers are poorly repaired⁵⁴. In TCR, the ability of a lesion (whether of the NER- or BER-type) to block RNA polymerase seems critical (stage I in the figure opposite). The stalled polymerase must be displaced to make the injury accessible for repair⁵⁵, and this requires at least two TCR-specific factors: CSB and CSA. The subsequent stages of GG-NER and TCR may be identical. The XPB and XPD helicases of the multi-subunit transcription factor TFIIH open ~30 base pairs of DNA around the damage (II). XPA probably confirms the presence of damage by probing for abnormal backbone structure⁵⁶, and when absent aborts NER⁵³. The single-stranded-binding protein RPA (replication protein A) stabilizes the open intermediate by binding to the undamaged strand (III). The use of subsequent factors, each with limited capacity for lesion detection *in toto*, still allows very high damage specificity⁵⁷. The endonuclease duo of the NER team, XPG and ERCC1/XPF, respectively cleave 3' and 5' of the borders of the opened stretch only in the damaged strand, generating a 24–32-base oligonucleotide containing the injury (IV). The regular DNA replication machinery then completes the repair by filling the gap (V). In total, 25 or more proteins participate in NER. *In vivo* studies indicate that the NER machinery is assembled in a step-wise fashion from individual components at the site of a lesion. After a single repair event (which takes several minutes) the entire complex is disassembled again⁵⁸.



exhibits a dramatic >1000-fold incidence of sun-induced skin cancer. Frequency of internal tumours is modestly elevated and accelerated neurodegeneration is often noted. The disorder arises from mutations in one of seven genes (*XPA–XPG*). Cockayne syndrome, caused by mutation in the *CSA* or *CSB* genes, is a TCR-specific disorder that is remarkably dissimilar from xeroderma pigmentosum. No predisposition to cancer is observed, which may be explained by the fact that the TCR defect causes Cockayne syndrome cells to be particularly sensitive to lesion-induced apoptosis, thereby protecting against tumorigenesis. Physical and neurological development are impaired, resulting in dwarfism and dysmyelination. The syndrome includes features of premature ageing, which may be related to the increased trigger for apoptosis induced by transcriptional arrest from endogenous lesions in

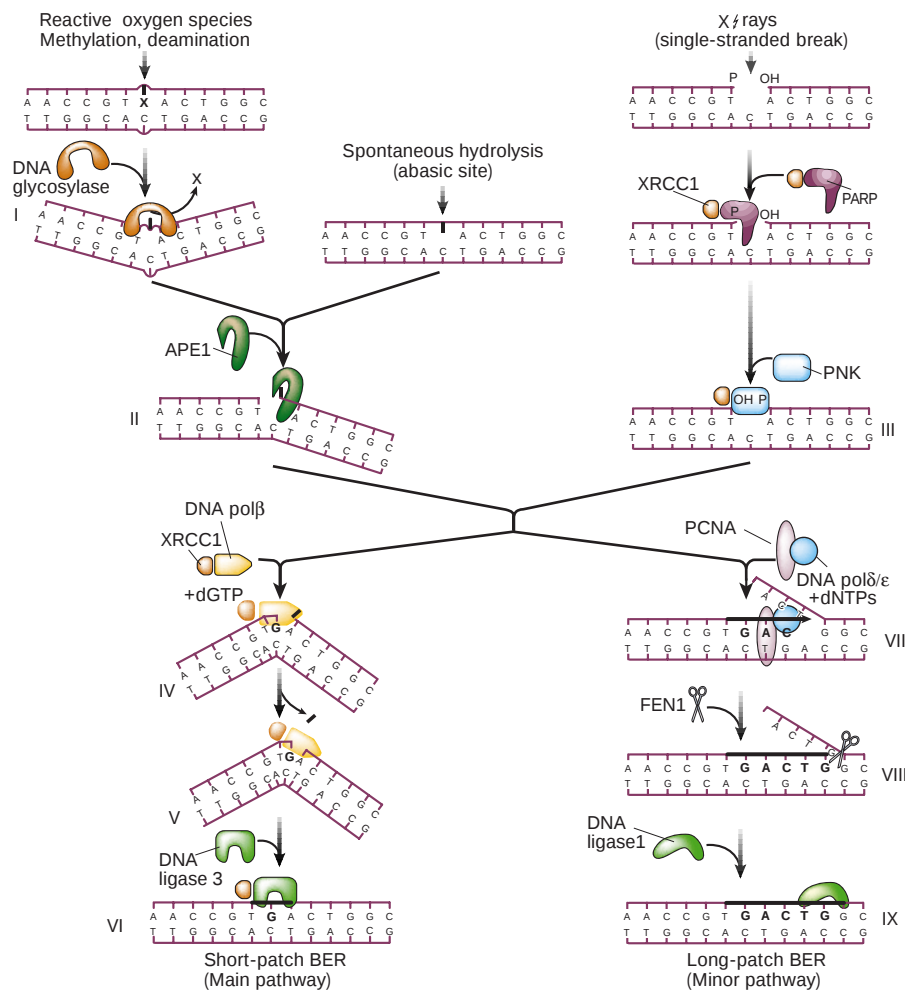
combination with the TCR defect. TTD is a condition sharing many symptoms with Cockayne syndrome, but with the additional hallmarks of brittle hair, nails and scaly skin. Mutations in the *XPD* or *XPB* genes can give rise to all three diseases. This puzzle is explained by the fact that, as subunits of TFIIH, XPD and XPB have dual functions: NER and transcription initiation. Mutations may not only compromise NER, but also affect transcription, causing developmental delay and reduced expression of the matrix proteins that causes brittle hair and scaly skin²⁰.

For almost all NER factors, mouse mutants have been generated²¹. Overall, the NER defect is accurately preserved, although cancer predisposition is more pronounced and neurological complications are milder in mice. Moreover, mice exhibit features of premature ageing.

Box 2

Mechanism for base-excision repair

A battery of glycosylases, each dealing with a relatively narrow, partially overlapping spectrum of lesions, feeds into a core reaction. Glycosylases flip the suspected base out of the helix by DNA backbone compression to accommodate it in an internal cavity of the protein. Inside the protein, the damaged base is cleaved from the sugar-phosphate backbone (stage I in the figure). The resulting abasic site can also occur spontaneously by hydrolysis. The core BER reaction is initiated by strand incision at the abasic site by the APE1 endonuclease (II). Poly(ADP-ribose) polymerase (PARP), which binds to and is activated by DNA strand breaks, and the recently identified polynucleotide kinase (PNK)⁵⁹ may be important when BER is initiated from a SSB to protect and trim the ends for repair synthesis (III). In mammals, the so-called short-patch repair is the dominant mode for the remainder of the reaction. DNA pol β performs a one-nucleotide gap-filling reaction (IV) and removes the 5'-terminal baseless sugar residue via its lyase activity (V); this is then followed by sealing of the remaining nick by the XRCC1-ligase3 complex (VI). The XRCC1 scaffold protein interacts with most of the above BER core components and may therefore be instrumental in protein exchange. The long-patch repair mode involves DNA pol β , pol δ/ϵ and proliferating cell nuclear antigen (PCNA) for repair synthesis (2–10 bases) as well as the FEN1 endonuclease to remove the displaced DNA flap and DNA ligase 1 for sealing (VII–IX). The above BER reaction operates across the genome. However, some BER lesions block transcription, and in this case the problem is dealt with by the TCR pathway described above, including TFIIH, XPG (which also stimulates some of the glycosylases) and probably the remainder of the core NER apparatus.



Base-excision repair

BER is the main guardian against damage due to cellular metabolism, including that resulting from reactive oxygen species, methylation, deamination and hydroxylation. The molecular mechanism¹³ has been resolved to the tertiary structure of all core components^{22–24} and is explained in Box 2.

BER and cancer

No human disorders caused by inherited BER deficiencies have been identified. Mouse models generated in recent years may provide an explanation: knockout of individual glycosylases does not cause an overt phenotype, which is explained by partial redundancy between different glycosylases^{13,25} and overlap with TCR. In fact, even a number of double mutants show only mild phenotypes, although mutagenesis and cancer susceptibility are probably increased. But inactivation of BER core proteins induces embryonic lethality, highlighting the vital importance of the process as a whole. This might be due to the contribution of spontaneously occurring abasic sites and SSBs that directly feed into the BER core reaction (Box 2) and/or to the generation of reaction intermediates by

the glycosylases that cannot be further processed^{13,25}. Interestingly, specific polymorphisms in XRCC1 seem associated with lung and other cancers²⁶.

DSB repair: homologous recombination and end joining

DSBs arise from ionizing radiation or X-rays, free radicals, chemicals and during replication of a SSB. After DSB detection, a complex cascade of reactions is triggered aimed at halting the cell-cycle machinery and recruiting repair factors^{11,27} (Fig. 5). One of the early initiators is the ataxia telangiectasia mutated (ATM) protein kinase, which is defective in the cancer-prone, X-ray-sensitive syndrome ataxia telangiectasia²⁸. Arrest in G1 is mediated via p53. Another early event, which depends on the giant protein-kinases ATM, ATR (ataxia telangiectasia related) and DNA-PK ϵ , is phosphorylation of histone H2AX in the DNA domain next to the DSB over a megadalton distance²⁹. This may provide a local chromatin state required for the complex repair reactions or for recruiting repair proteins. Homologous recombination and end joining are the main repair modes. When, after replication, a second identical DNA copy is available, homologous recombination seems to be preferred; otherwise cells

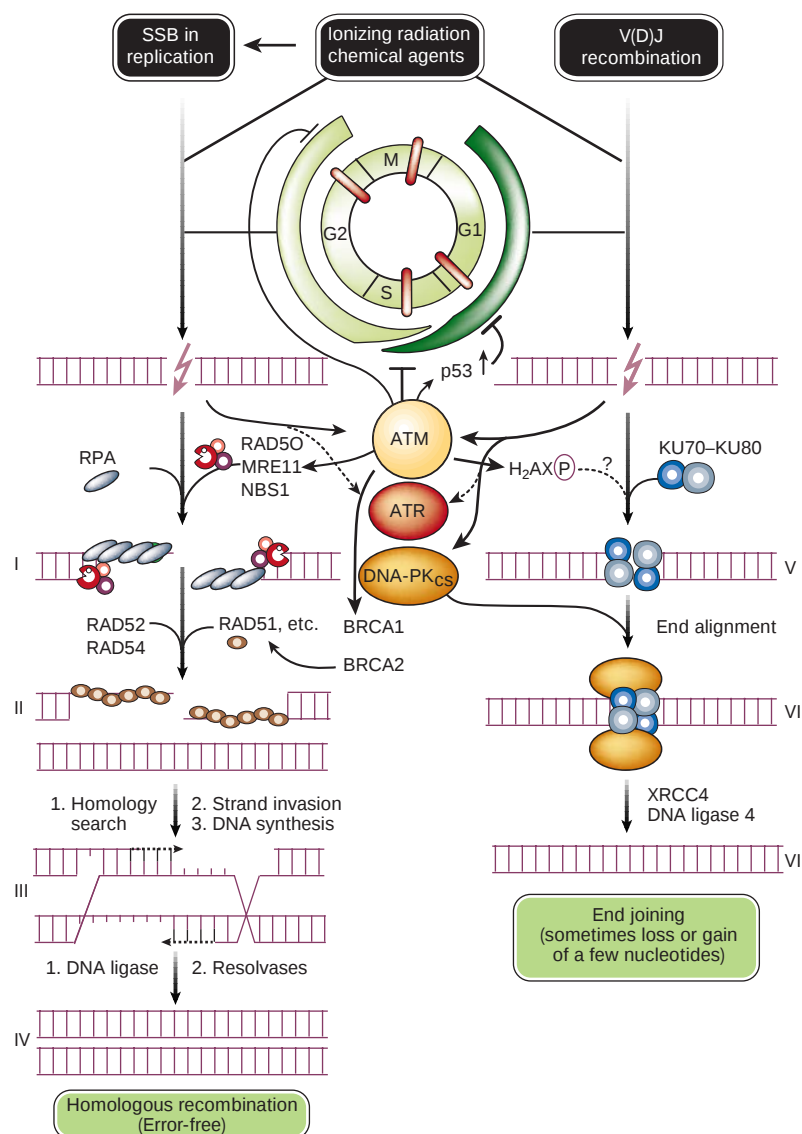
Box 3

Mechanism of homologous recombination and end joining

A tentative scenario for the homologous-recombination reaction is depicted in the left panel of the figure. To promote strand invasion into homologous sequences, the 5'–3' exonuclease activity of the RAD50/MRE11/NBS1 complex (also a substrate for ATM phosphorylation) exposes both 3' ends³⁰ (I). RPA facilitates assembly of a RAD51 nucleoprotein filament that probably includes RAD51-related proteins XRCC2, XRCC3, RAD51B, C and D. RAD52 stimulates filament assembly (II). RAD51 has, like its *Escherichia coli* RecA counterpart, the ability to exchange the single strand with the same sequence from a double-stranded DNA molecule. Correct positioning of the sister chromatids by cohesins probably facilitates the identification of a homologous sequence. A candidate for the complex chromatin transactions associated with these DNA gymnastics is RAD54, a member of the SWI/SNF family of DNA-dependent ATPases. After identification of the identical sister chromatid sequence, the intact double-stranded copy is used as a template to properly heal the broken ends by DNA synthesis (III). Finally, the so-called Holliday-junctions are resolved by resolvases^{27,33,60} (IV). Homologous recombination involves the simultaneous action of large numbers of the same molecules, which are found to be concentrated in radiation-induced nuclear foci. These depend on, and also include, the BRCA1 and BRCA2 proteins³⁶. Recent evidence implicates BRCA2 directly or indirectly in nuclear translocation of RAD51 (ref. 61).

Cells in G1 have only the homologous chromosome for recombination repair. However, this may be difficult to find in the complex genome. Moreover, it is potentially dangerous as a template for repair as it may lead to homozygosity for recessive mutations. As an alternative, the end-joining reaction simply links ends of a DSB together, without any template, using the end-binding KU70/80 complex and DNA-PK_{cs}, followed by ligation by XRCC4–ligase4 (reviewed by 27,33; see the right panel of the figure, stages V–VII). The function of KU70/80 might involve end protection and approximating the ends, in addition to a signalling function by DNA-PK_{cs}. End joining may be further facilitated when the ends are still held together through nucleosomes or other structures. End joining is sometimes associated with gain or loss of a few nucleotides if internal microhomologies are used for annealing before sealing. This implies the involvement of DNA polymerases and/or nucleases. Note that the KU complex is also involved in telomere metabolism^{27,62}.

Cells in G1 have only the homologous chromosome for recombination repair. However, this may be difficult to find in the complex genome. Moreover, it is potentially dangerous as a template for repair as it may lead to homozygosity for recessive mutations. As an alternative, the end-joining reaction simply links ends of a DSB together, without any template, using the end-binding KU70/80 complex and DNA-PK_{cs}, followed by ligation by XRCC4–ligase4 (reviewed by 27,33; see the right panel of the figure, stages V–VII). The function of KU70/80 might involve end protection and approximating the ends, in addition to a signalling function by DNA-PK_{cs}. End joining may be further facilitated when the ends are still held together through nucleosomes or other structures. End joining is sometimes associated with gain or loss of a few nucleotides if internal microhomologies are used for annealing before sealing. This implies the involvement of DNA polymerases and/or nucleases. Note that the KU complex is also involved in telomere metabolism^{27,62}.



rely on end joining, which is more error-prone. Their presumed mechanisms are explained in Box 3.

DSB repair and cancer

Besides ataxia telangiectasia, mutations in *MRE11* give rise to an ataxia telangiectasia-like disorder, whereas defects in *NBS1* are associated with the Nijmegen breakage syndrome (NBS)³⁰ (Table 1). All three conditions display cancer predisposition (particularly lymphomas), immunodeficiency, hypersensitivity to X-rays and chromosomal instability. Ataxia telangiectasia is additionally characterized by ataxia, cerebellar degeneration and ocular telangiectasia, whereas the cardinal symptoms of NBS are microcephaly and growth retardation^{28,31}. Inherited defects in *BRCA1* and *BRCA2* strongly predispose to breast cancer. In addition, cancer-prone

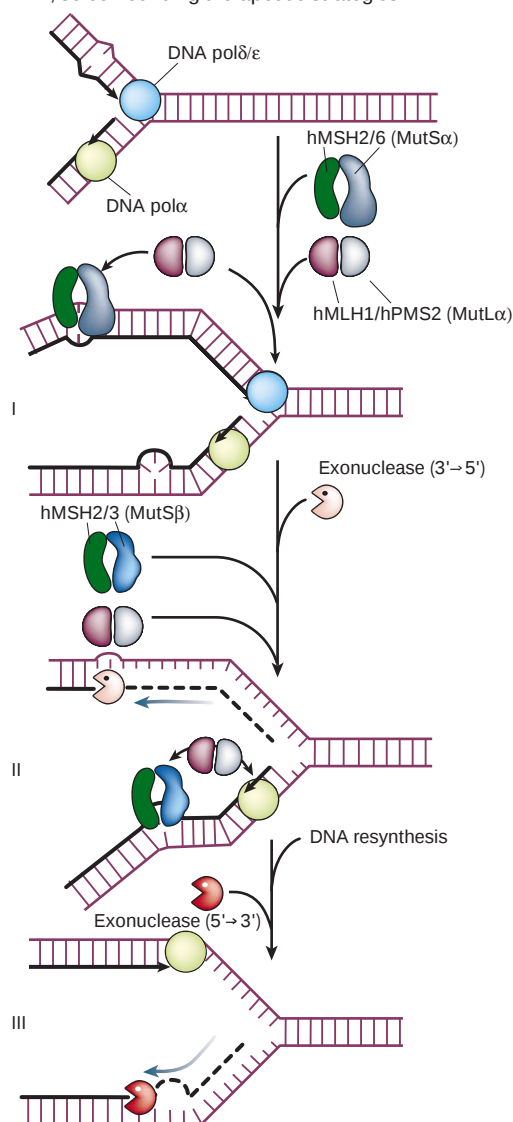
chromosomal-instability conditions such as Werner, Bloom and Rothmund Thomson syndrome, which all involve RecQ-like helicases, might carry defects in homologous recombination (Table 1). Inborn defects in the ligase IV component of end joining have been described for a single patient with leukaemia³².

Except for ATM, mice with null mutations in the above homologous-recombination factors tend to suffer from early embryonic lethality or in some cases display a mild phenotype (Rad52, Rad54), presumably because of functional redundancy³³. Lethality is preceded by gross chromosomal rearrangements, perhaps because endogenous lesions such as SSBs are converted to DSBs upon replication. The viable phenotype of mice and patients with ataxia telangiectasia may be due to partial functional overlap with ATR and DNA-PK_{cs}. Recently, double-mutant mice of *ATM* and *DNAPK_{cs}* were indeed

Box 4

Model for mismatch repair

Mammalian MMR involves multi-member families of the *E. coli* prototype factors MutS and MutL^{63,64}. Heterodimers of hMSH2/6 (called hMutS α) focus on mismatches and single-base loops (stage I in the figure below, upper strand), whereas hMSH2/3 dimers (hMutS β) recognize insertion/deletion loops (II, lower strand). Heterodimeric complexes of the hMutL-like proteins hMLH1/hPMS2 (hMutL α) and hMLH1/hPMS1 (hMutL β) interact with MSH complexes and replication factors. Strand discrimination may be based on contact with the nearby replication machinery. A number of proteins are implicated in the excision of the new strand past the mismatch and resynthesis steps, including pol δ/ϵ , RPA, PCNA, RFC, exonuclease 1, and endonuclease FEN1 (II, III). MMR components also interact functionally with NER and recombination. Recent crystallographic studies have revealed that a MutS dimer detects the structural instability of a heteroduplex by kinking the DNA at the site of the mismatch, which is facilitated when base pairing is affected^{65,66}. However, DNA damage with similar characteristics, such as that caused by alkylating agents and intercalators, may fool MutS, triggering erroneous or futile MMR. Intact MMR thus confers sensitivity, and as several of these agents are used in chemotherapy, tumours may become resistant to them on the basis of selection for defective MMR, so confounding therapeutic strategies⁶⁷.



found to be lethal³⁴. Inactivation of *ATR* by itself is inviable already at the blastocyst stage. Inactivation of *BRCA1* and *BRCA2* in mice is also embryonically lethal; cell lines display defects in homologous recombination^{35–37}.

The severe phenotype of the mouse mutants and the highly cancer-prone human syndromes highlight the importance of homologous recombination. Mouse KU mutants display sensitivity to agents that lead to breaks in DNA, and have immunological problems because the KU proteins are involved in V(D)J recombination of antibody gene sequences. In addition, these mutants display poor development, several features of premature ageing and increased apoptosis of postmitotic neurons in the developing brain. Mice with defects in DNA-PK ϵ (SCID mice) display a similar but generally milder phenotype. In contrast, XRCC4- and ligase IV-knockout mice seem more severe, with late embryonic lethality resulting from massive ATM- and p53-dependent neuronal apoptosis^{33,38}.

Mismatch repair

Specific sequence motifs comprised of dinucleotide repeats are unstable in some human cancers³⁹. This phenotype of 'microsatellite instability' is caused by defects in MMR in the hereditary non-polyposis colorectal cancer (HNPCC) and in a variety of sporadic cancers. MMR removes nucleotides mispaired by DNA polymerases and insertion/deletion loops (ranging from one to ten or more bases) that result from slippage during replication of repetitive sequences or during recombination. Defects in this system dramatically increase mutation rates, fuelling the process of oncogenesis. Four principal steps in MMR can be delineated: (1) mismatch recognition; (2) recruitment of additional MMR factors; (3) search for a signal that identifies the wrong (newly synthesized) strand, followed by degradation past the mismatch; and (4) resynthesis of the excised tract. A tentative model is depicted in Box 4.

MMR and cancer

Germline mutations in *hMLH1* and *hMSH2* together account for approximately half of all HNPCC patients, with *hMLH1* being responsible for most (~60%) of these cases. Defects in *hMSH6* cause late-onset atypical HNPCC. No *hMSH3* mutations have been reported. This is consistent with the notion that loss of *hMLH1* and *hMSH2* is associated with complete inactivation of MMR, whereas defects in the other proteins causes only a partial MMR deficiency. Mutations in *hPMS2* and *hPMS1* have been reported only in very few cases⁴⁰, implying that other factors have still to be identified. The reason why these MMR defects cause predominantly cancers of the colon, endometrium and ovary is still unclear.

Surprisingly, homozygous MMR deficiencies in mice are compatible with normal (albeit cancer-prone) development⁴¹. Mutants exhibit the expected molecular defects in terms of mutagenesis based on the role of the corresponding protein in MMR. Null mutations in the key genes *Mlh1* and *Msh2* predispose the mice mainly to lymphomas, although gastrointestinal tumorigenesis is also enhanced. This phenotype is similar to the combined *Msh3/6* defect, whereas a single *Msh3* or *Msh6* mutation induces cancer at a later age. *Pms2*^{-/-} mice display mainly haematological malignancies, but no intestinal neoplasias. In addition, *Mlh1*^{-/-} males are sterile owing to the occurrence of apoptosis during meiosis; this occurs secondary to the premature separation of chromosomes, which suggests a role of MLH in meiotic recombination. Null alleles of the MutS homologues *Msh4* and *Msh5* display infertility for both sexes, indicating unique functions of these genes in gametogenesis.

The telomere-division limiter

Telomeres constitute the caps of chromosome ends, and function as a buffer to prevent loss of important genomic sequence during replication. In humans they consist of a 5–15-kilobase repeated array of the sequence TTAGGG bound by a specific set of proteins. DNA replication proceeding in the 5'→3' direction needs an RNA primer

before it can initiate. Therefore, it leaves a terminal stretch of unreplicated DNA at the 5'-end of linear molecules. This leads to loss of a number of the terminal telomeric repeats with every S-phase, shortening telomeres by about 100 base pairs per cell division. In the germ line and in some specific tissues, telomere length is maintained by a specialized reverse transcriptase, called telomerase, adding new repeats using a tightly associated RNA template to compensate for the loss (reviewed in ref. 42). However, in many human cells and tissues telomerase activity is low or absent⁴³, leading to gradual telomere attrition with each cell division. This limits the replicative capacity of a cell but also prevents the outgrowth of a transformed cell to a full-blown tumour⁴⁴.

Telomeres in human tumours are often shorter compared with the tissue from which they derive. A large proportion (>90%) of cancer cells has reacquired telomerase activity at a shorter set length⁴³, demonstrating the need for an active telomere metabolism to sustain tumour growth^{45,46}. Furthermore, late-generation mice lacking functional telomerase seem to be resistant to skin carcinogenesis⁴⁷, indicating that the telomere-division limitator is relevant for preventing this type of epithelial carcinomas. However, it has been proposed that when telomeres become too short, a transient period of genomic havoc is induced^{48,49}. This stage of chromosomal instability could fuel the degeneration of the tumour to a more malignant state, for example via a loss of checkpoint functions leading to a scenario in which chromosomal damage is tolerated and actually drives tumour evolution. This stage of massive genomic instability could contribute to reacquisition of the telomerase activity⁴⁸ or to invention of alternative mechanisms that can solve the telomere problem⁵⁰. This may explain the increase in the incidence of spontaneous tumours in highly proliferative cell types such as lymphomas and teratocarcinomas, which are apparent in late-generation telomerase-deficient mouse mutants⁵¹. Thus, the consequences of the telomere-division limitation may have both advantages and disadvantages in terms of carcinogenesis^{48,49}.

Concluding remarks and perspectives

Research over the past few years has provided ample evidence that genome instability is one of the main forces driving the onset and progression of carcinogenesis. Genetic degeneration is linked intimately with all aspects of maintenance of DNA integrity and gene function and is fuelled by the continuous erosion of the genome by environmental and endogenous genotoxic agents. The outlines of key systems involved are rapidly emerging, although we may still be missing other important mechanisms, including error-free damage tolerance, additional intricacies of homologous recombination, and chromatin-modification mechanisms. For instance, *de novo* hypermethylation of CpG-rich islands nested in gene promoters seems to be a common means of silencing tumour-suppressor genes in cancer⁵². The era of postgenomics will enable the delineation of the complex response of cells, tissues and intact organisms against DNA injury, disclosing the intricate interactions between DNA repair, replication, transcription, chromatin dynamics, cell-cycle progression and apoptosis. The study of DNA maintenance mechanisms will not only reveal the biological impact of the havoc time wreaks on the genome, including oncogenesis and age-related diseases, but should also uncover new paradigms for prevention, genetic susceptibility, diagnosis and rational therapy. □

- Loeb, L. A. Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res.* **51**, 3075–3079 (1991).
- Cadet, J., Berger, M., Douki, T. & Ravanat, J. L. Oxidative damage to DNA: formation, measurement, and biological significance. *Rev. Physiol. Biochem. Pharmacol.* **131**, 1–87 (1997).
- Finkel, T. & Holbrook, N. J. Oxidants, oxidative stress and the biology of ageing. *Nature* **408**, 239–247 (2000).
- Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
- Yamaizumi, M. & Sugano, T. UV-induced nuclear accumulation of p53 is evoked through DNA damage of actively transcribed genes independent of the cell cycle. *Oncogene* **9**, 2775–2784 (1994).
- Goodman, M. E. & Toppin, B. S. Sloppier copier DNA polymerases involved in genome repair. *Curr. Opin. Genet. Dev.* **10**, 162–168 (2000).
- Kunkel, T. A. & Bebenek, K. DNA replication fidelity. *Annu. Rev. Biochem.* **69**, 497–529 (2000).

- Lawrence, C. The Rad6 DNA repair pathway in *Saccharomyces cerevisiae*: what does it do and how does it do it? *BioEssays* **16**, 253–258 (1994).
- Masutani, C. *et al.* The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase η . *Nature* **399**, 700–704 (1999).
- Johnson, R. E., Kondratieff, C. M., Prakash, S. & Prakash, L. hRAD30 mutations in the variant form of xeroderma pigmentosum. *Science* **285**, 263–265 (1999).
- Zhou, B. B. & Elledge, S. J. The DNA damage response: putting checkpoints in perspective. *Nature* **408**, 433–439 (2000).
- Friedberg, E. C., Walker, G. C. & Siede, W. *DNA Repair and Mutagenesis* (ASM Press, Washington DC, 1995).
- Lindahl, T. & Wood, R. D. Quality control by DNA repair. *Science* **286**, 1897–1905 (1999).
- Takata, M. *et al.* Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *EMBO J.* **17**, 5497–5508 (1998).
- Tornaletti, S. & Hanawalt, P. C. Effect of DNA lesions on transcription elongation. *Biochimie* **81**, 139–148 (1999).
- de Laat, W. L., Jaspers, N. G. & Hoeijmakers, J. H. Molecular mechanism of nucleotide excision repair. *Genes Dev.* **13**, 768–785 (1999).
- Batty, D. P. & Wood, R. D. Damage recognition in nucleotide excision repair of DNA. *Gene* **241**, 193–204 (2000).
- Bootsma, D., Kraemer, K. H., Cleaver, J. & Hoeijmakers, J. H. J. in *The Metabolic and Molecular Basis of Inherited Disease* Vol. 1 (eds Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D.) 677–703 (McGraw-Hill, New York, 2001).
- Lehmann, A. R. The xeroderma pigmentosum group D (XPD) gene: one gene, two functions, three diseases. *Genes Dev.* **15**, 15–23 (2001).
- Vermeulen, W. *et al.* A temperature-sensitive disorder in basal transcription and DNA repair in humans. *Nature Genet.* **27**, 299–303 (2001).
- de Boer, J. & Hoeijmakers, J. H. Cancer from the outside, aging from the inside: mouse models to study the consequences of defective nucleotide excision repair. *Biochimie* **81**, 127–137 (1999).
- Wyatt, M. D., Allan, J. M., Lau, A. Y., Ellenberger, T. E. & Samson, L. D. 3-methyladenine DNA glycosylases: structure, function, and biological importance. *BioEssays* **21**, 668–676 (1999).
- Mol, C. D., Parikh, S. S., Putnam, C. D., Lo, T. P. & Tainer, J. A. DNA repair mechanisms for the recognition and removal of damaged DNA bases. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 101–128 (1999).
- Bruner, S. D., Norman, D. P. & Verdine, G. L. Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. *Nature* **403**, 859–866 (2000).
- Wilson, D. M. & Thompson, L. H. Life without DNA repair. *Proc. Natl Acad. Sci. USA* **94**, 12754–12757 (1997).
- Divine, K. K. *et al.* The XRCC1 399 glutamine allele is a risk factor for adenocarcinoma of the lung. *Mutat. Res.* **461**, 273–278 (2001).
- Khanna, K. K. & Jackson, S. P. DNA double-strand breaks: signaling, repair and the cancer connection. *Nature Genet.* **27**, 247–254 (2001).
- Rotman, G. & Shiloh, Y. ATM: from gene to function. *Hum. Mol. Genet.* **7**, 1555–1563 (1998).
- Paull, T. T. *et al.* A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage. *Curr. Biol.* **10**, 886–895 (2000).
- Petrini, J. H. The Mre11 complex and ATM: collaborating to navigate S phase. *Curr. Opin. Cell Biol.* **12**, 293–296 (2000).
- Digweed, M., Reis, A. & Sperling, K. Nijmegen breakage syndrome: consequences of defective DNA double strand break repair. *BioEssays* **21**, 649–656 (1999).
- Ribalto, E. *et al.* Identification of a defect in DNA ligase IV in a radiosensitive leukaemia patient. *Curr. Biol.* **9**, 699–702 (1999).
- Van Gent, D. C., Hoeijmakers, J. H. J. & Kanaar, R. Chromosomal stability and the DNA double-strand break connection. *Nature Rev. Genet.* **2**, 196–206 (2001).
- Gurley, K. E. & Kemp, C. J. Synthetic lethality between mutation in Atm and DNA-PKcs during murine embryogenesis. *Curr. Biol.* **11**, 191–194 (2001).
- Jasin, M. Chromosome breaks and genomic instability. *Cancer Invest.* **18**, 78–86 (2000).
- Scully, R. & Livingston, D. M. In search of the tumour-suppressor functions of BRCA1 and BRCA2. *Nature* **408**, 429–432 (2000).
- Moynahan, M. E., Pierce, A. J. & Jasin, M. BRCA2 is required for homology-directed repair of chromosomal breaks. *Mol. Cell* **7**, 263–272 (2001).
- Gao, Y. *et al.* Interplay of p53 and DNA-repair protein XRCC4 in tumorigenesis, genomic stability and development. *Nature* **404**, 897–900 (2000).
- Jiricny, J. Replication errors: cha(lle)nging the genome. *EMBO J.* **17**, 6427–6436 (1998).
- Jiricny, J. & Nystrom-Lahti, M. Mismatch repair defects in cancer. *Curr. Opin. Genet. Dev.* **10**, 157–161 (2000).
- Heyer, J., Yang, K., Lipkin, M., Edelman, W. & Kucherlapati, R. Mouse models for colorectal cancer. *Oncogene* **18**, 5325–5333 (1999).
- McEachern, M. J., Krauskopf, A. & Blackburn, E. H. Telomeres and their control. *Annu. Rev. Genet.* **34**, 331–358 (2000).
- Shay, J. W. & Bacchetti, S. A survey of telomerase activity in human cancer. *Eur. J. Cancer* **33**, 787–791 (1997).
- Counter, C. M. *et al.* Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *EMBO J.* **11**, 1921–1929 (1992).
- Greider, C. W. Telomerase activity, cell proliferation, and cancer. *Proc. Natl Acad. Sci. USA* **95**, 90–92 (1998).
- Tang, R., Cheng, A. J., Wang, J. Y. & Wang, T. C. Close correlation between telomerase expression and adenomatous polyp progression in multistep colorectal carcinogenesis. *Cancer Res.* **58**, 4052–4054 (1998).
- Gonzalez-Suarez, E., Samper, E., Flores, J. M. & Blasco, M. A. Telomerase-deficient mice with short telomeres are resistant to skin tumorigenesis. *Nature Genet.* **26**, 114–117 (2000).
- DePinho, R. A. The age of cancer. *Nature* **408**, 248–254 (2000).
- de Lange, T. & Jacks, T. For better or worse? Telomerase inhibition and cancer. *Cell* **98**, 273–275 (1999).
- Bryan, T. M., Englezou, A., Dalla-Pozza, L., Dunham, M. A. & Reddel, R. R. Evidence for an alternative mechanism for maintaining telomere length in human tumors and tumor-derived cell lines. *Nature Med.* **3**, 1271–1274 (1997).

51. Rudolph, K. L. *et al.* Longevity, stress response, and cancer in aging telomerase-deficient mice. *Cell* **96**, 701–712 (1999).
52. Jones, P. A. & Laird, P. W. Cancer epigenetics comes of age. *Nature Genet.* **21**, 163–167 (1999).
53. Sugawara, K. *et al.* A multistep damage recognition mechanism for global genomic nucleotide excision repair. *Genes Dev.* **15**, 507–521 (2001).
54. Tang, J. Y., Hwang, B. J., Ford, J. M., Hanawalt, P. C. & Chu, G. Xeroderma pigmentosum p48 gene enhances global genomic repair and suppresses UV-induced mutagenesis. *Mol. Cell* **5**, 737–744 (2000).
55. Le Page, F. *et al.* Transcription-coupled repair of 8-oxoguanine: requirement for XPG, TFIIH, and CSB and implications for Cockayne syndrome. *Cell* **101**, 159–171 (2000).
56. Buschta-Hedayat, N., Buterin, T., Hess, M. T., Missura, M. & Naegeli, H. Recognition of nonhybridizing base pairs during nucleotide excision repair of DNA. *Proc. Natl Acad. Sci. USA* **96**, 6090–6095 (1999).
57. Sugawara, K. *et al.* Xeroderma pigmentosum group C protein complex is the initiator of global genome nucleotide excision repair. *Mol. Cell* **2**, 223–232 (1998).
58. Houtsmuller, A. B. *et al.* Action of DNA repair endonuclease ERCC1/XPF in living cells. *Science* **284**, 958–961 (1999).
59. Whitehouse, C. J. *et al.* XRCC1 stimulates human polynucleotide kinase activity at damaged DNA termini and accelerates DNA single-strand break repair. *Cell* **104**, 107–117 (2001).
60. Haber, J. E. Partners and pathways repairing a double-strand break. *Trends Genet.* **16**, 259–264 (2000).
61. Davies, A. A. *et al.* Role of BRCA2 in control of RAD51 recombination and DNA repair protein. *Mol. Cell* **7**, 273–282 (2001).
62. Hsu, H. L. *et al.* Ku acts in a unique way at the mammalian telomere to prevent end joining. *Genes Dev.* **14**, 2807–2812 (2000).
63. Kolodner, R. D. & Marsischky, G. T. Eukaryotic DNA mismatch repair. *Curr. Opin. Genet. Dev.* **9**, 89–96 (1999).
64. Harfe, B. D. & Jinks-Robertson, S. DNA mismatch repair and genetic instability. *Annu. Rev. Genet.* **34**, 359–399 (2000).
65. Obmolova, G., Ban, C., Hsieh, P. & Yang, W. Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature* **407**, 703–710 (2000).
66. Lamers, M. H. *et al.* The crystal structure of DNA mismatch repair protein MutS binding to a G·T mismatch. *Nature* **407**, 711–717 (2000).
67. Karran, P. & Bignami, M. DNA damage tolerance, mismatch repair and genome instability. *BioEssays* **16**, 833–839 (1994).

Acknowledgements

I apologize to my colleagues for being able to cite only recent original papers and for the use of reviews for earlier relevant work, owing to space and reference limitations. I thank L. Niedernhofer for valuable help with the manuscript. J.H.J.H. is supported by the Dutch Cancer Society, the Dutch Science Organization (medical and chemical divisions), the EC, NIH and IACR.

The microenvironment of the tumour–host interface

Lance A. Liotta & Elise C. Kohn

Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland 20892, USA

Throughout the entire process of cancer aetiology, progression and metastasis, the microenvironment of the local host tissue can be an active participant. Invasion occurs within a tumour–host microecology, where stroma and tumour cells exchange enzymes and cytokines that modify the local extracellular matrix, stimulate migration, and promote proliferation and survival. A new class of cancer therapies that targets this pathological communication interface between tumour cells and host cells is currently under development.

Neoplasia can be considered a pathological imbalance of tissue-cell societies^{1–3}. Malignancy is a state that emerges from a tumour–host microenvironment⁴ in which the host participates in the induction, selection and expansion of the neoplastic cells. Rather than being renegades, malignant tumour cells recruit vasculature and stroma through production and secretion of stimulatory growth factors and cytokines³. The locally activated host microenvironment (both cellular and extracellular elements) in turn modifies the proliferative and invasive behaviour of the tumour cells^{5,6}.

Cancer invasion can be viewed as a derangement in the proper sorting of cell populations, causing a violation of normal tissue boundaries. Tissue architecture, maintained normally by basement-membrane delineation of tissue boundaries and cell–cell communication, suppresses inappropriate intermixing of cells from different tissue types. Cells remain confined to their home territory because they are held in check by intercommunication with neighbouring cells and with the surrounding extracellular matrix (ECM). Appropriate sorting of parenchymal tissue cells during morphogenesis and wound healing^{2,3} may be tightly regulated by soluble- and solid-phase stimuli. In contrast, successful malignant tumour cells can be hypothesized as being resistant to the regulatory signals because they may appropriate, misinterpret or disregard these signals and dominate the local invaded host-cell populations⁷. Indeed, investigators have identified specific molecules comprising tumour–host signal pathways that promote cancer progression.

Cross-talk between mesenchyme and epithelium has been described as a known driver of differentiation and development, with recent examples in prostate and ovary^{5,8}. Furthermore, studies have shown that changes in stromal behaviour can promote epithelial transformation^{5,9,10}. For example, a well characterized interaction between mesenchyme and epithelium mediates the cyclical regulation of the ovary before, and during, the reproductive period⁸. It is possible that, during menopausal changes, the loss of this interactive milieu may be involved in the promotion of epithelial ovarian cancer. Thus, derangements in the normal cell–cell conversation that occurs during embryogenesis and postnatal development may set the stage for cancer aetiology^{2,10}. Altered cell–cell and cell–substratum survival signals may release normal constraints, thereby enabling malignant cells to migrate across tissue boundaries^{1,11}.

Physiological and malignant invasion use similar molecular mechanisms; the difference is that malignant invasion persists¹¹. Neovascularization, wound healing, and neurite outgrowth during embryogenesis are examples of physiological invasion. In response to trophic signals, vascular cells, wounded epithelial sheets or neurites will migrate, penetrate tissue barriers and establish appropriate new anastomoses^{12–14}. But when the trophic signal is removed or the injury is repaired, physiological invasion ceases. Malignant cells perpetually stimulate host stromal and vascular cells to conduct physiological invasion. Within the same microenvironment, vascular sprouts migrate and invade towards the tumour mass while tumour cells migrate outwards in the opposite direction^{12–14}. Activation of the local invasive environment seems to create a permissive field for the malignant cell^{2–4} (Fig. 1).

Who is invading whom?

The interaction between the epithelial and mesenchymal compartments creates a local heterotypic ‘invasion field’ from which the metastatic cell emerges and disseminates. It is unclear who is invading whom (Figs 1, 2). The transition from normal to invasive carcinoma is preceded by, or is concomitant with, activation of local host stroma¹⁵. For example, disorganization and disruption of the periglandular basement membrane and hemidesmosome structure¹⁵ in the breast is concomitant with a local neovascular ‘blush’ and has been shown to precede³ frank malignant conversion during the transition from *in situ* to invasive carcinoma^{16,17}. A bed of permeable vessels can be identified beneath ovarian cancer spheroids at the time of implantation¹⁸.

After traversal of the basement membrane, the ECM on the stromal side of the epithelial basement membrane is the next site of tumour–host interactions. The pericellular ECM of the epithelial and endothelial compartments differs from the stromal matrix and has been shown to influence epithelial cell function in both malignant behaviour and nonmalignant differentiation^{2,5,9,19}. Stromal-cell activation may be reflected in modifications of the adjacent ECM that are favourable to the microinvasion of cancer cells³. A variety of cell types populate the stromal compartment, ranging from immune cells (for example, lymphocytes and dendritic cells; see article in this issue by Rosenberg, pages 380–384), inflammatory cells (for example, monocytes and granulocytes), muscle and myofibroblasts, and vascular cells (for example, lympho-endothelial, vascular-endothelial cells and pericytes)². The

relative abundance of each cell type may change at the local site of tumour cell invasion¹⁷. It is unclear whether the changes in stromal cell population occur before or after the invasion takes place^{2,5,10,20}. During metastasis, tumour cells leave the stroma and enter nearby lymphatic and blood vascular channels. Intimate tumour–endothelial adhesive interactions occur at the site where tumour cells traverse the vessel wall, and junctions between vascular cells are retracted. Intravascular circulating tumour cells attach to endothelium in the target organ and are stimulated to grow as colonies inside the vessel²¹. These observations support the concept that tumour cells in one sense are actually captured by the vascular cells during entry and exit from the circulation.

Remodelling of the ECM, which is confined to the immediate pericellular environment of the cell, seems to be a necessary step in local invasion^{11,22}. The principal enzymes that degrade the ECM and cell-associated proteins are: (1) the matrix metalloproteinases (MMPs), a family of secreted and membrane-anchored proteinases; (2) the adamalysin-related membrane proteinases; (3) the bone morphogenetic protein-1-type metalloproteinases; and (4) tissue serine proteinases, including tissue plasminogen activator, urokinase, thrombin and plasmin²². This complement of enzyme classes is tightly and exquisitely regulated by a series of activation steps and specific inhibitors. In a striking demonstration of host–tumour interdependence, most of the enzymes and inhibitors complexed at the invasion front are contributed by host cells, not by the invading tumour cells^{20,23–25} (Fig. 2).

The enzyme cascade is confined to the cell surface at the point of invading pseudopodia by binding the enzymes to adhesion sites, cell-surface receptors and adjacent ECM molecules^{24–26}. MT1-MMP, an ECM-degrading enzyme, contains a transmembrane/cytoplasmic sequence that confines it to microinvasion sites on the surface of the tumour-cell invadopodia (Fig. 2). In complex with one of the tissue inhibitors of metalloproteinases (TIMP-2), it becomes a receptor and activator of MMP-2 (ref. 25), a soluble MMP produced by stromal fibroblasts and endothelial cells. The serine proteinase uPA (for urokinase plasminogen activator) is confined to the invading pseudopodia through a cooperation between integrins and the uPA receptor (uPAR)²⁷. uPAR is an adhesion receptor for vitronectin, and also interacts laterally with integrin β -chains. Proteolysis of ECM proteins modifies integrin-mediated anchorage, focal adhesions and cytoskeletal architecture, and triggers signalling molecules such as focal adhesion kinase (FAK)^{28,29}. Such heterotypic complexes direct and confine the enzymatic field to the forward edge of the invading cell, leaving intact the peripheral and distal attachment sites required for traction. As the invading cell moves forward through ECM barriers, the leading-edge complex of enzymes, inhibitors and receptor molecules cycle through adhesion, de-adhesion and proteolysis. The direction of tumour-cell invasion and migration can be influenced by chemoattractants and by the construction of preferred adhesion pathways. Local attractants include scatter factor/hepatocyte growth factor (SF/HGF), which binds to the Met receptor (c-Met)^{4,30}; proteolysed matrix fragments, which are recognized by integrins³¹; or cytokines and growth factors, such as epidermal growth factor (EGF) and transforming growth factor (TGF)- β released from the degraded matrix³². Cryptic Arg-Gly-Asp (RGD) sites exposed by proteolysis^{31,33–35} may guide the path in front of invading host or tumour cells (Fig. 2).

The organ preference for metastatic colonization is influenced heavily by communication between the circulating tumour cell and the target host tissue. Chemokines are growth factor-like molecules that bind to G-protein-coupled receptors. Circulating leukocytes and stem cells are known to use chemokine mechanisms to home in on specific organs³⁶. They induce leukocytes to adhere tightly to endothelial cells, and migrate towards the highest concentration of chemokine. As this behaviour seemed identical to that required for metastatic tumour cells, Muller *et al.*³⁷ hypothesized that tumour cells may co-opt the same chemokines to direct metastatic organ

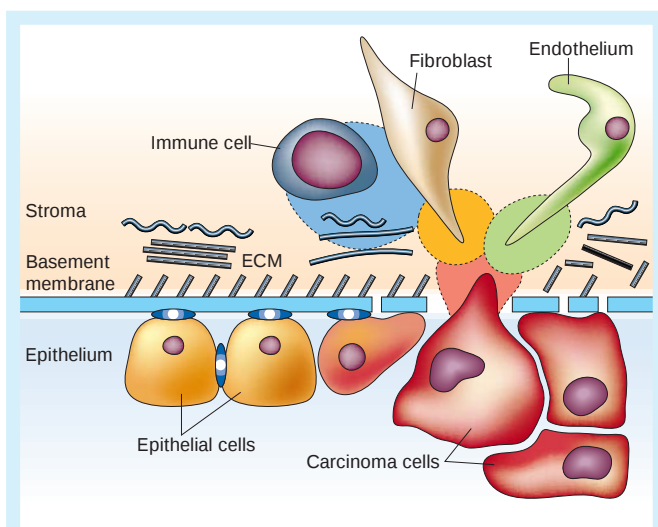


Figure 1 Microecology of the tumour–host invasion field. Invasive carcinoma is viewed as a pathology of multiple cell societies inhabiting the epithelial/mesenchymal stromal unit. Transition to invasive carcinoma is preceded by the activation of host fibroblasts, immune cells and endothelial cells. Invasion occurs in a localized zone of cross-talk and cooperation between the stromal cells and the premalignant epithelium (depicted as zones demarked by dashed lines). Cytokine and enzyme exchange between the participating cells stimulates migration of both cell types towards each other and modifies the adjacent extracellular matrix/basement membrane. The result is a breakdown of normal tissue boundaries.

preference. They conducted a comprehensive survey of known chemokines and found a receptor/ligand pair (CCR4 and CXCL12) that fit the profile expected for breast cancer metastasis homing to bone, lung and liver. *In vitro*, the CXCL12 ligand stimulated breast cancer cells to carry out the basics of invasion, including pseudopodial protrusion, directed migration and penetration of ECM barriers. *In vivo*, metastasis to CXCL12-rich lung tissue was blocked in animal models by treatment with a neutralizing anti-human CXCR4 monoclonal antibody³⁷.

Molecular cross-talk links motility, survival and proliferation

At the molecular level, cytoskeletal remodelling, adhesion and de-adhesion are not only required for cellular motility, but also are linked to proliferation and pro-survival pathways (Fig. 2). Integrins exist at the communication juncture between the cell and the ECM. Disengagement of integrin-mediated adhesion to the matrix, which is required for cellular translocation, can trigger apoptosis (programmed cell death) or anoikis^{38,39} if not followed by attachment and readhesion¹³. Consequently, for a cell to invade into, and migrate within, a mesenchymal environment, pro-invasive and antianoikis signals must occur in concert.

Pro-invasive and pro-survival messages (Fig. 2) converge at numerous pathway nodes⁴⁰, diverging again into multiple effector molecules. Integrin engagement activates multiple downstream molecules necessary for motile function and survival. FAK, whose phosphorylation is necessary for functional focal adhesion signalling and migration^{29,41}, was shown to be an early component of the pro-survival pathways combating anoikis^{38,39}. FAK phosphorylation also links integrin-mediated signals to the Ras/mitogen-activated protein kinase (MAPK-ERK) pathway^{28,29,42}. In addition, HGF binding to the Met receptor stimulates the tyrosine phosphorylation of FAK and its association with the signal-transducing adaptor protein Grb2, thus connecting Met to the Ras pathway⁴³. The Met receptor may also be associated with the EGF receptor³⁰. Many signalling molecules are involved in regulating cell motility, including myosin light chain kinase (MLCK)⁴⁴, β -catenin, FAK, phosphatidylinositol 3-OH kinase (PI(3)K), Ras, Rac,

Rho and Cdc42. Thus, there is significant overlap between the motility and invasion pathways and those driving survival of cancer and stromal cells.

Use of arrays to study the tumour–host microecology

Inferences about tumour–host interactions can be made using cell lines and animal models. However, these approaches may not accurately model human epithelial malignancies, which can evolve over a period of 5–15 years. Biopsy samples of human carcinoma often contain adjacent premalignant and nonmalignant precursors. Consequently, an individual patient's tissue sample may contain a 'memory' of the sequence of events that culminated over time in the malignant carcinoma. Gene expression microarrays and serial analysis of gene expression (SAGE) are promising technologies for directly analysing tissue^{45,46}, but currently they require a large amount of input material, and so cannot be easily used to study the microscopic populations of adjacent tumour cells and stroma.

The combination of microdissection and protein microarrays has been applied successfully to the microecology of early-stage cancer. Protein lysate microarrays consist of very small (picogram) quantities of protein lysates from cell lines, whole lysed tissue, or microdissected subpopulations of lysed tissue cells, immobilized and arrayed on a solid phase⁴⁷. The array can be probed with antibodies recognizing phosphorylated forms of signal proteins. Detection is highly sensitive, quantitative and precise, so that the state of signal pathways can be profiled. Individual subpopulations of host and tumour tissue cells within a microscopic field of invasion or premalignant transition can be microdissected and studied individually. Using protein microarray technology, lysate microarrays showed that activation of PI(3)K substrates, and suppression of apoptosis, are early events in the microenvironment of prostate cancer evolution. Proteomic analysis of tissue lysate arrays provided direct quantitative evidence that suppression of apoptosis in human prostate intraepithelial neoplasia and invasive prostate cancer is associated with phosphorylation of the serine/threonine kinase Akt (a

downstream target of PI(3)K) and suppression of downstream apoptotic caspases. In addition, protein arrays applied to the cancer cells at the invasion front revealed a pattern of phosphorylation of extracellular signal-regulated protein kinase (ERK)²⁸, supporting an uncoupling of signals from integrins and growth factors⁴⁷.

Stromal therapy emerges as a new strategy

If we view the cancer state as a product of its microenvironment, and can identify the molecular signals that participate in tumour–host cross-talk, then stromal therapy could emerge as a viable approach to cancer prevention and intervention. Stromal therapy addresses an early but dynamic target and can therefore be applied at multiple points in the treatment cascade, from primary chemoprevention to treatment for relapsed and disseminated disease. Low-dose stromal therapy might potentially reverse subtle, but critical, imbalances in tumour–host signals, and possess a specificity that could minimize collateral toxicity to uninvolved tissues.

It is possible that some of the signal-regulatory and anti-invasive agents already under development might be useful anti-stromal agents^{12,14,48}. The most promising cellular target for anti-invasion treatment may be the stromal fibroblasts and endothelial cells (see Table 1 for selected example targets and agents).

Several agents currently under investigation have been shown to disrupt cell adhesion, or the downstream signals propagated through integrins. CAI (carboxyamido-triazole), a modulator of transmembrane calcium uptake, was shown to regulate phosphorylation of FAK, now known to be important in transmitting migration, adhesion and matrix-survival signals^{41,49,50}. The nonsteroidal anti-inflammatory agents sulindac sulphide and caffeic acid phenethyl ester (a phenolic antioxidant) regulate integrin-mediated signalling pathways⁵¹. The last two agents, when used in subapoptotic doses, were able to cause rearrangement of the actin cytoskeleton with loss of focal adhesion plaques. FAK phosphorylation was reduced and the abrogation of FAK signalling resulted in a reduction in invasive capacity.

Figure 2 Molecular cross-talk at the invasion front. Example mediators are shown. Motility and invasion is a bi-directional process. Fibroblasts produce chemoattractants such as SF/HGF, which stimulates motility of tumour cells by binding to the Met receptor (c-Met). Tumour cells produce angiogenesis factors such as VEGF and bFGF, which bind to receptors on stromal vascular cells and cause increased vascular permeability, endothelial proliferation, migration and invasion. Fibroblasts and endothelial stromal cells elaborate latent enzymes, including MMPs and uPA, which dock on the surface of the carcinoma invadopodia and become activated, thereby degrading the ECM, and clearing a pathway. ECM degradation releases bound growth factors such as TGF- β and EGF, which bind to cognate receptors (TGF- β R and uPAR) on the carcinoma cell. ECM proteolysis also exposes cryptic RGD sites, which are recognized by integrins. Cross-talk between signal pathways within the carcinoma cells links motility, proliferation and pro-survival signals. For example, phosphorylation of FAK through Met and integrin signalling transduces signals through Ras, PI(3)K, β -catenin and MLCK, causing cytoskeletal remodelling, ERK activation of mitogenesis, and sustainment of survival through phosphorylation of Akt.

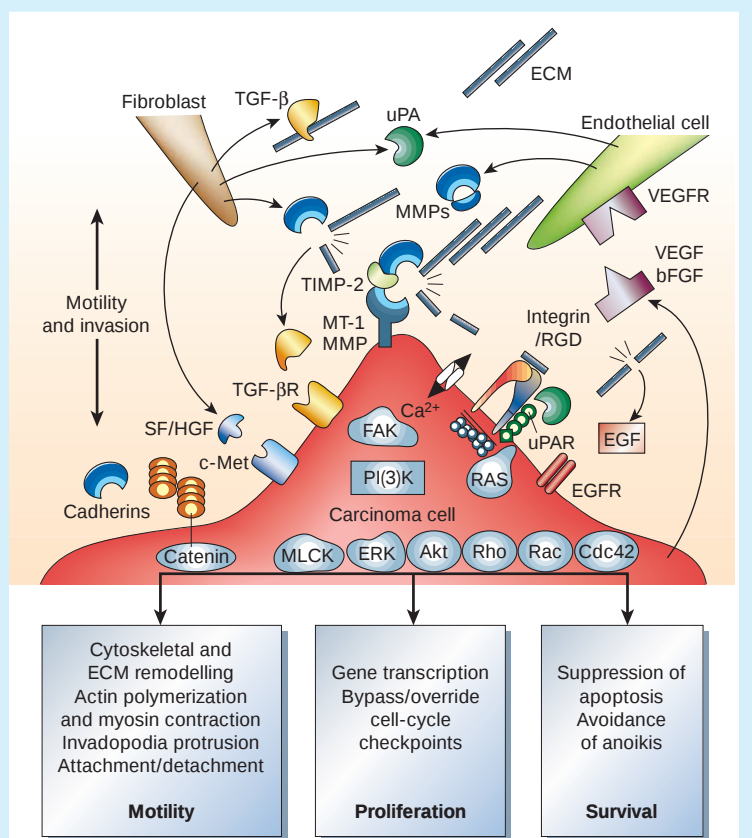


Table 1 Therapeutics targeting stroma–tumour interactions

Target	Example agent	Comments
Adhesion/attachment	RGD-toxin constructs and RGD-targeted gene therapy	Have not reached clinical trials
	Anti- $\alpha v \beta 3$ monoclonal antibody (Vitaxin, Medi522)	Cytostasis in patients; anti-tumour and anti-angiogenic in animal models
Proteolysis	Matrix metalloproteinase inhibitors	Cytostatic in patients; rare occurrence of tumour partial regressions; stromal fibrosis; activity seen in multiple animal models and in combination with chemotherapy; new agents with varied MMP specificity under development
Motility	No selective agents	
Signal pathways	Squalamine (NHE-3 inhibitor)	Selective to endothelial cells
	PDGFR, KDR and EGFR small molecule inhibitors	Active <i>in vitro</i> in animal models; preclinical activity in combinations; phase I trials completed of several agents, some tumour stabilization or regression
	Anti-EGFR monoclonal antibody (C225)	Neutralizing antibody; active <i>in vitro</i> in animal models; phase I trials ongoing
	Anti-VEGF antibody	Blocking antibody; active <i>in vitro</i> in animal models; preclinical activity alone and in combination; phase I–III trials ongoing
	CAI (nonvoltage-gated Ca^{2+} uptake inhibitor)	Active <i>in vitro</i> in animal models; preclinical activity in combinations; phase I trials of single agents and combinations, some tumour stabilization or regression
Extracellular matrix	Pirfenidone	Suppresses stromal/inflammatory cell remodelling by stroma expression of TGF- β . Phase I trials for pulmonary fibrosis

Additional agents currently under clinical study that might double as stromal therapies fall into several categories: (1) enzyme and protease inhibitors⁵², including MMP inhibitors⁵³; (2) anti-adhesive molecules, such as anti-integrin peptides or antibodies^{48,54–56}; (3) signal modulators^{57,58}, including ion-flux blockers⁵⁹ and inhibitors of the tyrosine kinase pathway^{60,61}; and (4) antifibrotic drugs such as pirfenidone⁶². Integrin-targeting agents, such as antibodies or peptides that block integrin $\alpha v \beta 3$, do alter motile and survival functions of responsive stromal and endothelial cells^{48,54–56}. Pirfenidone, used as an experimental agent to suppress bleomycin-induced pulmonary fibrosis, has been shown to reduce the influx of activated macrophages and inflammatory cells and to downregulate the overexpression of TGF- β , event that precede the ECM changes associated with fibrosis⁶². Pirfenidone, therefore, is a candidate agent that can be used to study the stromal cell populations within the tumour–host microenvironment.

Extracellular growth factor and cytokine ligands constitute targets for stromal therapy. Ongoing strategies for inhibiting tumour angiogenesis are aimed at blocking extracellular angiogenesis factors (for example, vascular endothelial growth factor (VEGF) and basic fibroblast growth factor (bFGF)) which stimulate vascular permeability, growth and stromal invasion. New extracellular mediators are being identified that offer fresh approaches to stromal therapy. Production and secretion of lysophosphatidic acid (LPA) by epithelial ovarian cancer cells is an example of a lipid mediator found in the extracellular space of ovarian cancer tissue⁶³. LPA binds to cognate receptors on tumour cells and host cells, resulting in activation of the PI(3)K/Akt pathway and promoting increased genomic expression of the p110 catalytic subunit⁶⁴. Molecular modulation of the expression of one of the LPA receptors, edg-2, within ovarian cancer cells was shown to stimulate apoptosis and anoikis in an

LPA-independent fashion⁶⁵. LPA signalling may promote the survival and dissemination of ovarian cancer cells within the peritoneal cavity.

Concluding remarks

The malignant state is unleashed by defects in communication pathways which recruit host cells to become active participants in the heterotypic tissue invasion field. Cross-talk between tumour cells and a variety of host cell types triggers pro-survival, proliferation and invasion pathways in both the cancer cells and their host. Future developments will include a new class of therapies targeting the extracellular and intracellular mediators that act at the tumour–host communication interface. □

- Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
- Park, C. C., Bissell, M. J. & Barcellos-Hoff, M. H. The influence of the microenvironment on the malignant phenotype. *Mol. Med. Today* **6**, 324–329 (2000).
- Brown, L. F. *et al.* Vascular stroma formation in carcinoma in situ, invasive carcinoma, and metastatic carcinoma of the breast. *Clin. Cancer Res.* **5**, 1041–1056 (1999).
- Wernert, N. The multiple roles of tumor stroma. *Virchows Arch.* **430**, 433–443 (1997).
- Aboiseif, S., El-Sakka, A., Young, P. & Cunha, G. Mesenchymal reprogramming of adult human epithelial differentiation. *Differentiation* **65**, 113–118 (1999).
- Vaccariello, M., Javaherian, A., Wang, Y., Fusenig, N. E. & Garlick, J. A. Cell interactions control the fate of malignant keratinocytes in an organotypic model of early neoplasia. *J. Invest. Dermatol.* **113**, 384–391 (1999).
- Fidler, I. Critical determinants of cancer metastasis: rationale for therapy. *Cancer Chemother. Pharmacol.* **43**, 3–10 (1999).
- Nilsson, E. & Skinner, M. K. Cellular interactions that control primordial follicle development and folliculogenesis. *J. Soc. Gynecol. Invest.* **8**, S17–S20 (2001).
- Li, Y., Liu, W., Hayward, S. W., Cunha, G. & Baskin, L. S. Plasticity of the urothelial phenotype: effects of gastro-intestinal mesenchyme/stroma and implications for urinary tract reconstruction. *Differentiation* **66**, 126–135 (2000).
- Olumi, A. F. *et al.* Carcinoma-associated fibroblasts direct tumor progression of initiated human prostatic epithelium. *Cancer Res.* **59**, 5002–5011 (1999).
- Liotta, L. A., Steeg, P. S. & Stetler-Stevenson, W. G. Cancer metastasis and angiogenesis: an imbalance of positive and negative regulation. *Cell* **64**, 327 (1991).
- Kohn, E. C. & Liotta, L. A. Molecular insights into cancer invasion: strategies for prevention and intervention. *Cancer Res.* **55**, 1856–1862 (1995).
- Stromblad, S. & Cheresh, D. A. Integrins, angiogenesis and vascular cell survival. *Chem. Biol.* **3**, 881–885 (1996).
- Carmeliet, P. & Jain, R. K. Angiogenesis in cancer and other diseases. *Nature* **407**, 249–2571 (2000).
- Tomakidi, P. *et al.* Defects in the basement membrane and hemidesmosomes structure correlate with malignant phenotype and stromal interactions in HaCaT-Ras xenografts. *Differentiation* **64**, 263–275 (1999).
- Guidi, A. J. *et al.* Vascular permeability factor (vascular endothelial growth factor) expression and angiogenesis in patients with ductal carcinoma in situ of the breast. *Cancer* **80**, 1945–1953 (1997).
- Lee, A. H., Dublin, E. A. & Bobrow, L. G. Angiogenesis and expression of thymidine phosphorylase by inflammatory and carcinoma cells in ductal carcinoma in situ of the breast. *J. Pathol.* **187**, 285–290 (1999).
- Gilead, A. & Neeman, M. Dynamic remodeling of the vascular bed precedes tumor growth: MSL ovarian carcinoma spheroids implanted in nude mice. *Neoplasia* **1**, 226–230 (1999).
- Airola, K. & Fusenig, N. E. Differential stromal regulation of MMP-1 expression in benign and malignant keratinocytes. *J. Invest. Dermatol.* **116**, 85–92 (2000).
- Sternlicht, M. D. *et al.* The stromal proteinase MMP-3/stromelysin-1 promotes mammary carcinogenesis. *Cell* **98**, 137–146 (1999).
- Mehdi, A. *et al.* Intravascular origin of metastasis from the proliferation of endothelium-attached tumor cells: a new model for metastasis. *Nature Med.* **6**, 100–106 (2000).
- Werb, Z. ECM and cell surface proteolysis: regulating cellular ecology. *Cell* **91**, 439–442 (1997).
- Coussens, L. M., Tinkle, C. L., Hanahan, D. & Werb, Z. MMP-9 supplied by bone marrow-derived cells contributes to skin carcinogenesis. *Cell* **103**, 481–490 (2000).
- Bowden, E., Barth, M., Thomas, D., Glazer, R. & Mueller, S. An invasion related complex of cortactin, paxillin, and PKC ζ associates with invadopodia at sites of extracellular matrix degradation. *Oncogene* **18**, 4440–4449 (1999).
- Nakahara, H. *et al.* Transmembrane/cytoplasmic domain mediated membrane type 1–matrix metalloproteinase docking to invadopodia is required for cell invasion. *Proc. Natl Acad. Sci. USA* **94**, 7959–7964 (1997).
- Hoegs, S., Oh, H.-R., Corcoran, M. & Stetler-Stevenson, W. Tissue inhibitor of metalloproteinases-2 (TIMP-2) suppresses TKR-growth factor signaling independent of metalloproteinase inhibition. *J. Biol. Chem.* **276**, 3203–3214 (2001).
- Andreasen, P., Kioller, L., Christensen, L. & Duffy, M. The urokinase type plasminogen activator system in cancer metastasis: a review. *Int. J. Cancer* **72**, 1–22 (1997).
- Fashena, S. & Thomas, S. M. Signaling by adhesion receptors. *Nature Cell Biol.* **2**, E225–E229 (2000).
- Braga, V. The crossroads between cell-cell adhesion and motility. *Nature Cell Biol.* **2**, E182–E184 (2000).
- Jo, M. *et al.* Cross talk between epidermal growth factor receptor and c-Met signal pathways in transformed cells. *J. Biol. Chem.* **275**, 8806–8811 (2000).
- Varner, J. A. & Cheresh, D. A. Integrins and cancer. *Curr. Opin. Cell Biol.* **8**, 724–730 (1996).
- Roberts, A. B., McCune, B. K. & Sporn, M. B. TGF- β : regulation of extracellular matrix. *Kidney Int.* **41**, 557–559 (1992).
- Varner, J. A., Brooks, P. C. & Cheresh, D. A. The integrin $\alpha v \beta 3$: angiogenesis and apoptosis. *Cell Adhes. Comm.* **3**, 367–374 (1995).
- Fukai, F. *et al.* Release of biologically activities from quiescent fibronectin by conformational change and limited proteolysis by matrix metalloproteinases. *Biochemistry* **34**, 11453–11459 (1995).

35. Davis, G. E. Affinity of integrins for damaged extracellular matrix: alpha v beta 3 binds to denatured collagen type I through RGD sites. *Biochem. Biophys. Res. Commun.* **182**, 1025–1031 (1992).
36. Zlotnik, A. & Yoshie, O. Chemokines: a new classification system and their role in immunity. *Immunity* **12**, 121–127 (2000).
37. Muller, A. *et al.* Involvement of chemokine receptors in breast cancer metastasis. *Nature* **410**, 50–56 (2001).
38. Frisch, S. M., Vuori, K., Ruoslahti, E. & Chan-Hui, P. Y. Control of adhesion-dependent cell survival by focal adhesion kinase. *J. Cell Biol.* **134**, 793–799 (1996).
39. Frisch, S. M. & Ruoslahti, E. Integrins and anoikis. *Curr. Opin. Cell Biol.* **9**, 701–706 (1997).
40. Cantley, L. & Pawson, T. Cell regulation. *Curr. Opin. Cell Biol.* **13**, 121–124 (2001).
41. Dusko, I. *et al.* Reduced cell motility and enhanced focal adhesion contact formation in cells from FAK-deficient mice. *Nature* **377**, 539–544 (1995).
42. Xue, L., Murray, J. H. & Tolkovsky, A. M. The Ras/phosphatidylinositol 3-kinase and Ras/ERK pathways function as independent survival modules each of which inhibits a distinct apoptotic signaling pathway in sympathetic neurons. *J. Biol. Chem.* **275**, 8817–8824 (2000).
43. Chen, H., Chan, P., Tang, M., Cheng, C. & Chang, T. Tyrosine phosphorylation of focal adhesion kinase stimulated by hepatocyte growth factor leads to mitogen-activated protein kinase activation. *J. Biol. Chem.* **273**, 25777–25782 (1998).
44. Totsukawa, G. *et al.* Distinct roles of ROCK and MLCK in spatial regulation of MLC phosphorylation for assembly of stress fibers and focal adhesions in 3T3 fibroblasts. *J. Cell Biol.* **150**, 797–806 (2000).
45. Liotta, L. A. & Petricoin, E. F. Molecular profiling of human cancer. *Nature Genet. Rev.* **1**, 48–56 (2000).
46. St. Croix, B. *et al.* Genes expressed in human tumor endothelium. *Science* **289**, 1197–1202 (2000).
47. Pawelczak, C. P. *et al.* Reverse phase protein microarrays which capture disease progression shown activation of pro-survival pathways at the cancer invasion front. *Oncogene* **20**, 1981–1989 (2001).
48. Gutheil, J. C. *et al.* Targeted antiangiogenic therapy for cancer using Vitaxin: a humanized monoclonal antibody to the integrin $\alpha_v\beta_3$. *Clin. Cancer Res.* **6**, 3056–3061 (2000).
49. Kohn, E. C. *et al.* Angiogenesis: role of calcium-mediated signal transduction. *Proc. Natl Acad. Sci. USA* **92**, 1307–1311 (1995).
50. Masiero, L., Lapidus, K. A., Ambudkar, I. & Kohn, E. C. Regulation of the RhoA pathway in human endothelial cell spreading on type IV collagen: role of calcium influx. *J. Cell. Sci.* **112**, 3205–3213 (1999).
51. Weyant, M. J., Carothers, A. M., Bertagnolli, M. E. & Bertagnolli, M. M. Colon cancer chemopreventive drugs modulate integrin-mediated signaling pathways. *Clin. Cancer Res.* **6**, 949–956 (2000).
52. Brown, P. D. & Giavazzi, R. Matrix metalloproteinase inhibition: a review of anti-tumour activity. *Annals Oncol.* **6**, 967–974 (1995).
53. Rowinsky, E. *et al.* Phase I and pharmacologic study of the specific matrix metalloproteinase inhibitor BAY 12-9566 on a protracted oral daily dosing schedule in patients with solid malignancies. *J. Clin. Oncol.* **18**, 178–186 (2000).
54. Brooks, P. C. *et al.* Anti-integrin alpha v beta 3 blocks human breast cancer growth and angiogenesis in human skin. *J. Clin. Invest.* **96**, 1815–1822 (1995).
55. Bonfoco, E., Chen, W., Paul, R., Cheres, D. A. & Cooper, N. R. $\beta 1$ integrin antagonism on adherent, differentiated human neuroblastoma cells triggers an apoptotic signaling pathway. *Neuroscience* **101**, 1145–1152 (2000).
56. Lode, H. N. *et al.* Synergy between an antiangiogenic integrin α_v antagonist and an antibody-cytokine fusion protein eradicates spontaneous tumor metastases. *Proc. Natl Acad. Sci. USA* **96**, 1591–1596 (1999).
57. Kohl, N. E. Farnesyltransferase inhibitors. Preclinical development. *Ann. NY Acad. Sci.* **886**, 91–102 (1999).
58. Sills, A. K. *et al.* Squalamine inhibits angiogenesis and solid tumor growth in vivo and perturbs embryonic vasculature. *Cancer Res.* **58**, 2784–2792 (1998).
59. Kohn, E. C. *et al.* A phase I trial of carboxyamido-triazole and paclitaxel for relapsed solid tumors: potential efficacy of the combination, and demonstration of pharmacokinetic interaction. *Clin. Cancer Res.* (in the press).
60. Yuan, Z. Q. *et al.* Frequent activation of AKT2 and induction of apoptosis by inhibition of phosphoinositide-3-OH kinase/Akt pathway in human ovarian cancer. *Oncogene* **19**, 2324–2330 (2000).
61. Perrotte, P. *et al.* Anti-epidermal growth factor receptor antibody C225 inhibits angiogenesis in human transitional cell carcinoma growing orthotopically in nude mice. *Clin. Cancer Res.* **5**, 257–265 (1999).
62. Iyer, S., Gurujeyalakshmi, G. & Giri, S. N. Effects of pirfenidone on transforming growth factor- β gene expression at the transcriptional level in bleomycin hamster model of lung fibrosis. *J. Pharmacol. Exp. Therapeut.* **291**, 367–373 (1999).
63. Fang, X. *et al.* Lysophospholipid growth factors in the initiation, progression, metastases and management of ovarian cancer. *Ann. NY Acad. Sci.* **905**, 188–208 (2000).
64. Shayesteh, L. *et al.* PIK3CA is implicated as an oncogene in ovarian cancer. *Nature Genet.* **21**, 99–102 (1999).
65. Furui, T. *et al.* Overexpression of edg-2/vzg-1 induces apoptosis and anoikis in ovarian cancer cells in a lysophosphatidic acid-independent manner. *Clin. Cancer Res.* **5**, 430–4318 (1999).

Progress in human tumour immunology and immunotherapy

Steven A. Rosenberg

Division of Clinical Sciences, National Cancer Institute, Building 10, Room 2B42, 10 Center Drive, MSC 1502, Bethesda, Maryland 20892-1502, USA

Studies of the administration of interleukin-2 to patients with metastatic melanoma or kidney cancer have shown that immunological manipulations can mediate the durable regression of metastatic cancer. The molecular identification of cancer antigens has opened new possibilities for the development of effective immunotherapies for patients with cancer. Clinical studies using immunization with peptides derived from cancer antigens have shown that high levels of lymphocytes with anti-tumour activity can be raised in cancer-bearing patients. Highly avid anti-tumour lymphocytes can be isolated from immunized patients and grown *in vitro* for use in cell-transfer therapies. Current studies are aimed at understanding the mechanisms that enable the cancer to escape from immune attack.

For much of the twentieth century, studies of the immunological response to tumours remained on the fringe of mainstream efforts in immunology. Scepticism was high concerning the existence of an immune response to cancer in humans and doubt existed concerning the applicability to humans of information derived from studies of transplantable murine tumours. A widely quoted article in the *British Journal of Cancer* reported no evidence of immune response to 27 different spontaneous tumours in mice and concluded that: "transplanted tumour systems...entail artifactual immunity associated with viral or chemical induction"¹. Another review commenting on cancer immunotherapy concluded that: "It would be as difficult to reject the right ear and leave the left ear intact as it is to immunize against cancer"².

Much has changed in the past 15 years, as increasing information about the molecular basis of tumour–host interactions has developed. The convergence of information resulting from basic studies in cellular immunology, along with increasing sophistication in biotechnology, which has made biologic reagents available in pharmacological amounts, has opened extraordinary possibilities for the development of effective immunotherapies for patients with cancer³. In addition, the ability to genetically modify cells involved in immunological reactions and to generate recombinant vectors containing genes encoding cancer antigens has resulted in early efforts at gene therapy of cancer.

During the past two decades, four sequential questions have characterized progress in the development of human cancer immunotherapy, discussion of which forms the basis of this review. Can immune manipulation cause the regression of established human cancers? What are the antigens involved in the immune recognition of human cancers? Can anti-tumour T cells be generated in patients by immunization with cancer antigens? What mechanisms limit cancer regression despite the *in vivo* generation of anti-tumour T cells?

Can immune manipulation cause cancer regression?

The first clear indication that immunological manipulations could cause the regression of established, invasive human cancers came from studies of the administration of interleukin-2 (IL-2) to humans with metastatic kidney cancer or melanoma⁴. IL-2, a cytokine produced by human

T-helper lymphocytes, has a panoply of immune regulatory effects, including the expansion of lymphocytes following activation by specific antigen. IL-2 has no direct impact on cancer cells, which can grow unimpeded *in vitro* in high concentrations of IL-2. Thus, the impact of IL-2 on cancers *in vivo* derives from its ability to expand lymphocytes with anti-tumour activity.

The administration of high-dose recombinant IL-2 to humans was reported to mediate the regression of even bulky, invasive tumours in selected patients with metastatic melanoma, kidney cancer and non-Hodgkin's lymphoma⁴. These initial studies showed that 15–20% of patients with these metastatic cancers sustained an objective cancer regression (50% total reduction), and complete regression of metastatic tumour occurred in half of these patients. In another study of 409 IL-2-treated patients, 8.1% of patients with metastatic melanoma or kidney cancer achieved a complete response and 9% achieved a partial response⁵. With a median follow-up of 7.1 years, 82% of these completely responding patients remained in continuous, ongoing, complete regression from 3 to over 12 years from the onset of treatment (Fig. 1), and many were probably cured. Studies of 255 patients with metastatic kidney cancer⁶ and 270 patients with metastatic melanoma⁷ from 22 different institutions achieved similar results. These studies showed that this relatively simple immunological manipulation could mediate the regression of human cancer in a variety of organs and spurred intensive efforts to understand, at a molecular level, these complex immunological anti-tumour events.

Which antigens are recognized in human cancers?

Multiple studies in experimental animals showed that cellular rather than humoral immune responses were responsible for the rejection of transplanted tumours or allogeneic (genetically different) tissues. With the exception of antibodies directed against growth factor receptors on cancer cells, the administration of antibodies has had little impact on the growth of solid tumours. Thus, significant effort has been devoted towards the identification of antigens recognized by human T lymphocytes^{8,9}.

Both CD8⁺ cytotoxic T cells and CD4⁺ T-helper cells recognize antigens presented as small peptides in the groove of surface human leukocyte antigen (HLA; the human analogue of the major histocompatibility complex (MHC))

molecules. CD8⁺ cells recognize peptides of 8–10 amino acids in length, derived from intracellular cytoplasmic proteins, digested in proteosomes and presented via the endoplasmic reticulum on cell-surface class I HLA molecules. In contrast, CD4⁺ cells use a different intracellular pathway and present engulfed extracellular proteins, digested to peptides in intracellular endosomes and presented on cell-surface class II HLA molecules. Thus, the recognition of antigens by T cells involves the recognition of both peptides and specific HLA molecules. These different pathways of antigen processing required the development of separate techniques to identify tumour antigens, but all depended on the ability to generate T lymphocytes capable of recognizing human cancer cells.

Many antigens recognized by CD8⁺ cells have been identified by transfecting complementary DNA libraries from tumour cells into target cells expressing the appropriate HLA molecule, and then using anti-tumour T cells to identify the appropriate transfectants^{8,9}. Alternatively, peptides eluted from the surface of human cancer cells (or from HLA molecules purified from cancer cells) can be pulsed onto antigen presenting cells (APCs) and tested for reactivity with specific anti-tumour lymphocytes^{10,11}. Purification and sequencing of these peptides can then lead to the identification of the parent protein.

A third technique often referred to as 'reverse immunology' has been used successfully to identify whether candidate proteins, selected because of their unique overexpression on cancer cells, represent cancer antigens¹². *In vitro* sensitization techniques are used to generate T cells that are reactive against the specific candidate antigens. If these T cells can also specifically recognize intact human cancer cells, the candidate protein is considered to be a tumour antigen. Another technique known as SEREX (serologic analysis of recombinant cDNA expression libraries)¹³ is based on the assumption that antibody production against a protein requires helper T cells. Diluted serum from cancer patients is used to detect proteins encoded by cancer cDNA libraries that are expressed in prokaryotes.

Because of the relative ease of generating human T cells that recognize melanomas, most human tumour antigens so far identified have been derived from this tumour type, although many antigens expressed on common epithelial tumours have also been identified. Examples of antigens recognized by CD8⁺ cells and presented on class I HLA molecules are presented in Table 1.

Knowledge of class II-restricted human cancer antigens recognized by CD4⁺ cells has lagged behind the identification of class I-restricted antigens. Transfection of cDNA libraries into target cells using common techniques is not effective because the encoded proteins do not travel to the class II pathway. But a new technique¹⁴ involving the screening of cDNA libraries fused to genes encoding invariant chain sequences designed to guide the transfected proteins into the class II presentation pathway has the potential for wide applicability. By transfecting these fusion vectors into APCs engineered to contain the appropriate molecules required for class II presentation, many new human tumour antigens recognized by CD4⁺ T cells have been identified. Examples of class II-restricted cancer antigens are presented in Table 1.

There is increasing evidence of a relationship between infectious agents and the incidence of cancer¹⁵. Many of the viruses associated with oncogenesis also present proteins on the induced cancers that can serve as targets for immune attack (Table 2). Thus, the E6 and E7 epitopes on cervical cancers caused by human papillomavirus, epitopes from Epstein–Barr virus (EBV) on lymphomas, and human T-cell lymphotropic virus-1 epitopes on adult T-cell leukaemias represent a different class of cancer antigens. Immunization against these antigens might be useful in cancer therapy, and elimination of these infectious agents might also be a strategy to help prevent cancer.

Many different intracellular proteins are known to represent human cancer antigens. Stoler *et al.* estimated that about 11,000 genomic alterations occur in a cancer cell, and such genomic instability provides multiple opportunities for the development of cancer antigens either by the overexpression of individual proteins or by the expression of mutated proteins¹⁶.

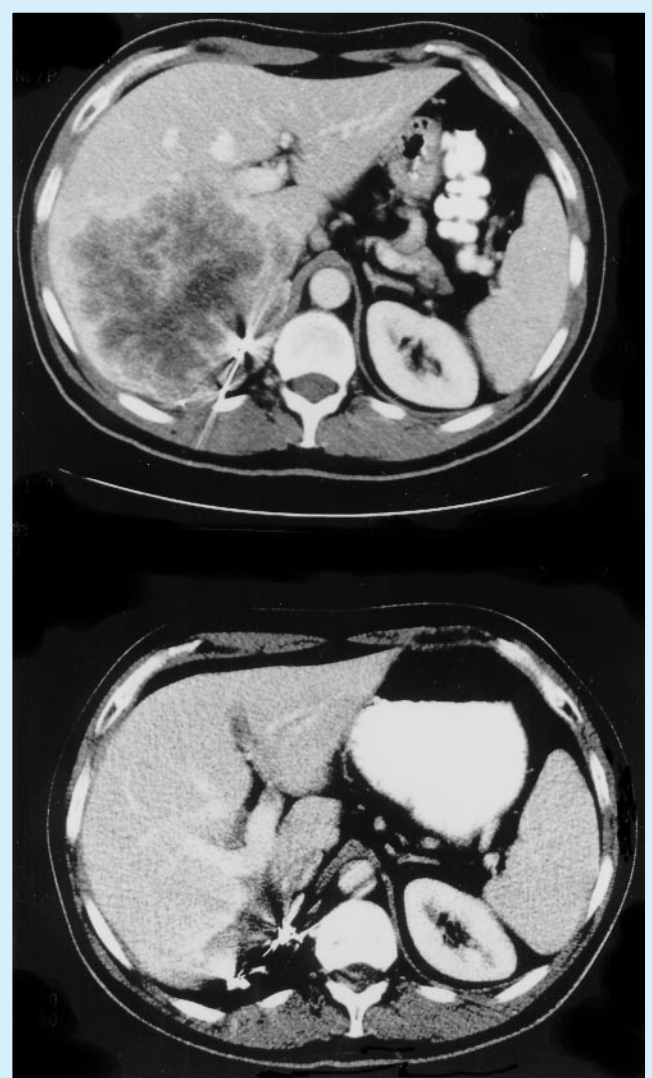


Figure 1 Complete regression of a large liver metastasis from kidney cancer in a patient treated with IL-2. Regression is ongoing seven years later.

Cancers of the haematopoietic system represent unique situations not shared by most cancers arising in solid tissues (the subject of this review). B lymphocytes can express unique idiotypes resulting from the gene rearrangements involved in antibody production. Because each B-cell clone gives rise to a lymphoma uniquely expressing this idiomorph, it can serve as a cancer antigen. The graft-versus-host reactions in patients with leukaemia undergoing allogeneic bone marrow transplantation can be associated with graft-versus-tumour effects that can enhance the therapeutic impact of chemotherapy (see article in this issue by Appelbaum, pages 385–389). The antigens that serve as targets of this immune attack have not been clearly identified.

Much has been learned in the past decade concerning cancer antigens on solid cancers. Four general principles from these findings are presented in Box 1.

Can immunization generate anti-tumour T cells?

Cancer immunotherapies fall into either active immunization or passive transfer approaches and the identification of cancer antigens has impacted on both areas³. Passive (or adoptive) approaches involve the transfer of immune cells with anti-tumour reactivity. Landsteiner and Chase first described the transfer of delayed hypersensitivity reactions from one animal to another using cells from sensitized donors¹⁷. Early studies of tumour immunity in mice showed that specific immunity to tumours could be transferred to normal mice using lymphocytes from

Table 1 Examples of human cancer antigens

Antigen	Reference	Antigen	Reference
I. Class I-restricted antigens recognized by CD8⁺ lymphocytes			
Melanoma-melanocyte differentiation antigens		Cancer-testes antigens	
MART-1 (Melan-A)	42	MAGE-1	48
gp100 (pmel-17)	43	MAGE-2	49
Tyrosinase	44	MAGE-3	50
Tyrosinase related protein-1	45	MAGE-12	51
Tyrosinase related protein-2	46	BAGE	52
Melanocyte-stimulating hormone receptor	47	GAGE	53
		NY-ESO-1	54, 55
Mutated antigens		Non-mutated shared antigens overexpressed on cancers	
β -catenin	56	α -Fetoprotein	62
MUM-1	57	Telomerase catalytic protein	63
CDK-4	58	G-250	64
Caspase-8	59	MUC-1	65
KIA 0205	60	Carcinoembryonic antigen	66
HLA-A2-R1701	61	p53	67
		Her-2/neu	68
II. Class II-restricted antigens recognized by CD4⁺ lymphocytes			
Epitopes from non-mutated proteins		Epitopes from mutated proteins	
gp100	69	Triosephosphate isomerase	73
MAGE-1	70	CDC-27	14
MAGE-3	70	LDLR-FUT	74
Tyrosinase	71		
NY-ESO-1	72		

Human cancer antigens restricted by HLA-A class I and recognized by CD8⁺ lymphocytes fall into four general categories. (1) Melanoma-melanocyte differentiation antigens are normal non-mutated proteins that are expressed exclusively on melanomas and on normal pigment-producing cells such as melanocytes. Lymphocytes that are reactive against these differentiation antigens can be found infiltrating into tumours. (2) Cancer-testes antigens can be widely expressed on a variety of epithelial tumours as well as on testis and placental tissue. (3) Mutated antigens represent normal proteins that contain mutations or translocations that give rise to unique epitopes. (4) Non-mutated shared antigens that are overexpressed on cancers. There is some evidence that overexpressed proteins, such as carcinoembryonic antigen, p53 and Her-2/neu, are tumour antigens, although evidence is controversial. As for antigens recognized by CD8⁺ cells, epitopes recognized by CD4⁺ cells are derived from both non-mutated and mutated proteins.

the spleen or peritoneal cavity of immunized donors^{18,19}. Early studies in humans, done before the identification of human cancer antigens, involved the transfer to tumour-bearing patients of lymphokine-activated killer (LAK) cells with non-HLA-restricted ability to recognize and lyse cancer cells *in vitro*²⁰. Despite the apparent success of LAK cells in treating micrometastases in experimental animals, clinical results in humans were disappointing. Techniques for growing large numbers of tumour-infiltrating lymphocytes (TILs) from resected tumours resulted in T-cell populations capable of specifically recognizing cancer antigens from about one-third of patients with melanoma^{21,22}. TILs could be expanded to 10¹⁰–10¹¹ cells and, when adoptively transferred along with IL-2 into melanoma patients, resulted in an objective response rate of about 35% (refs 23, 24). This objective regression rate was twice that seen with IL-2 alone and was also achieved in patients who had become refractory to treatment with IL-2 alone. In other studies²⁵, tumour regression resulted from adoptive transfer of either fresh or cultured donor lymphocytes in patients with lymphoproliferative disorders or lymphomas following allogeneic transplantation.

The ability to successfully immunize patients against defined cancer antigens has facilitated the *in vitro* generation of anti-tumour T cells that can be expanded and used for adoptive therapy²⁶. The ability to clone lymphocytes derived from single starting cells selected for their high avidity for tumour antigens, and to grow them to large numbers, is not only creating new possibilities for passive immunotherapy, but also provides a means of identifying the exact cellular characteristics that are required for mediating tumour rejection^{27–29}. The genetic modification of these lymphocytes to improve their anti-tumour efficacy (for instance, by inserting genes encoding anti-tumour or chemokine receptors or genes encoding anti-tumour cytokines) is under active investigation³⁰.

The achievements of active immunization against infectious diseases such as smallpox and polio have provided hope that cancer patients could be actively immunized against their own cancers to prevent or treat the disease. Before the identification of human cancer antigens, cancer vaccine approaches depended on immunization

Box 1

Principles of human cancer antigens

1. Cancer antigens can arise from: normal differentiation antigens; cancer-testes antigens; intronic sequences; alternative open reading frames; single-base mutations; post-transcriptional control of expression; chromosomal rearrangement; and aberrant processing.
2. A single cancer patient can develop immune reactions to multiple antigens, as shown by the reactivity of TILs. For example, TILs from patient 888 recognized: tyrosinase (differentiation antigen presented by HLA class I); β -catenin (class I mutation); P15 (class I antigen involved in post-transcriptional control); gp100 (class I intronic sequence; class II normal sequence); tyrosinase-related protein-1 (TRP-1; a class II differentiation antigen); TRP-2 (class II differentiation antigen); and Ki-67 (class II mutation). TILs from patient 586 recognized: TRP-1 (class I alternative open reading frame); TRP-2 (class I differentiation antigen); and NY-ESO-1 (cancer-testes antigen presented on class I (2 epitopes) and class II (2 epitopes) molecules).
3. A single cancer antigen contains epitopes that can be presented on many different surface HLA molecules. For example, the gp100 antigen is presented on HLA molecules A2, A3, A24, Cw8, DR4 and DR15, where the tyrosinase antigen is presented on A1, A2, A24, B44, DR4 and DR15.
4. Study of the immune reactivity of patients with melanoma can identify genes encoding antigens widely expressed on other tumours (such as breast and prostate cancers). These antigens can serve as targets for immune attack.

with either autologous or allogeneic whole cancer cells or cancer cell extracts (Box 2). But this approach is limited by the minute quantity of cancer antigenic molecules present in the intact cell. A variety of approaches to increase the immunogenicity of whole tumour cells has been attempted, including the injection of these cancer cells in a variety of adjuvants, or transducing cancer cells with genes encoding cytokines such as granulocyte-macrophage colony-stimulating factor, tumour necrosis factor or interferon- γ . Only limited evidence has been generated that these approaches can generate T cells in humans that can recognize intact tumour cells.

The identification of human cancer antigens has opened new approaches to the development of cancer vaccines (Box 2). Although often present in large amounts in the cell, epitopes from non-mutated differentiation antigens often exhibit low affinity for cell-surface HLA molecules. Mutated epitopes generally exhibit high affinity for HLA molecules, but often are derived from proteins with relatively poor expression in the cancer. Clinical trials using each of the different types of antigens will be required to determine which will be most effective in mediating anti-tumour immune effects. Multiple assays are available to assess the anti-tumour immune response of lymphocytes obtained before and after immunization. Assessment of immune status is often limited to circulating or lymph node lymphocytes rather than lymphocytes at the tumour site.

Recombinant expression of the genes encoding cancer antigens in *Escherichia coli*, yeast or baculovirus can result in the production of large quantities of purified cancer antigens for use in immunization, although the difficulty and expense of generating recombinant proteins that are suitable for human administration has significantly limited the application of this approach.

Many studies have used immunization with recombinant viruses that encode cancer antigens, including adenovirus, vaccinia virus and avipox^{31–34}. But only weak generation of anti-tumour T cells has been reported using these approaches, which is perhaps due to the presence of neutralizing antibodies that exist in most humans against the envelope proteins of these viral vectors. Many current studies emphasize the

Box 2

Vaccine approaches to cancer treatment

Vaccines can be based either on cancer cells or on the genetic identification of cancer antigens. Many of these materials can be used to pulse, transfect or transduce APCs or can be administered with a variety of adjuvants or cytokines.

Vaccines based on cancer cells are derived from: whole cancer cells (both autologous and allogeneic preparations); gene-modified cancer cells (genes encoding cytokines or co-stimulatory molecules); cancer cell extracts (lysates, membranes and heat-shock proteins); and cancer cells fused to APCs.

Vaccines based on the genetic identification of cancer antigens include: purified cancer antigens (natural or recombinant); synthetic peptides; 'naked' DNA (for example, plasmids); recombinant viruses (adenovirus, vaccinia or avipox); and recombinant bacteria (Bacille Calmete–Guérin and listeria).

use of recombinant avipox viruses, as humans have not previously been exposed to these viruses and the viruses cannot replicate in human tissue. To avoid possible immunization to viral envelope proteins, an alternate immunization approach has involved the direct injection of 'naked' DNA encoding cancer antigens into skin or muscle³⁵. The poor efficiency *in vivo* of transfection of DNA has limited its value for the generation of immune responses against cancer antigens, although successful immunization against infectious agents has been reported³⁶.

Increasing information concerning the importance of professional APCs such as dendritic cells or Langerhans cells in generating immune responses in humans has stimulated attempts to use these cells in cancer vaccines³⁷. These attempts have used APCs pulsed with recombinant tumour antigens, tumour lysates or tumour-derived peptides, or infected with recombinant viruses or RNA. More recently, immunization with dendritic cells fused to whole tumour cells has been reported³⁸.

T cells recognize peptides presented on the surface of tumour cells, a response that has led to immunization studies using immunodominant peptides derived from tumour antigens²⁶. This approach has been surprisingly successful for generating high levels of circulating T cells directed against cancer antigens. The immunogenicity of peptides derived from tumour antigens has been increased substantially by altering specific amino-acid residues at positions that anchor the peptide to the appropriate HLA molecule³⁹. Immunization with these modified peptides can result in as many as 4% of all circulating CD8⁺ T cells that are reactive with their own cancers. As the ability to immunize patients improves, the use of immunotherapy for the prevention of cancer recurrence in high-risk individuals represents an exciting area of clinical investigation.

When these approaches are used in the absence of cytokine administration, only sporadic instances of cancer regression result. Peptide vaccines given in conjunction with IL-2 may be capable of mediating substantially higher levels of cancer regression than administration of IL-2 alone. In one study, objective clinical responses were seen in 30–35% of patients receiving immunization with a modified peptide from the gp100 molecule (gp100:209-217(210M)) when administered with high-dose bolus IL-2 (refs 26, 40). This response rate was twice that seen in a large number of patients treated with a similar schedule of IL-2 alone². However, the simultaneous administration of peptide plus IL-2 resulted in a decrease in circulating anti-tumour cells compared to that found in patients receiving peptide alone, possibly due to traffic of specific lymphocytes to the tumour site, with a consequent decrease of these cells in the circulation.

What mechanisms limit cancer regression?

The identification in growing tumours of TILs with the ability to specifically recognize cancer antigens and destroy tumour cells *in vitro*, coupled with the ability to successfully immunize patients to

Table 2 Infections agents as antigens to prevent or treat cancer*

Bacteria	<i>Helicobacter pylori</i>	Gastric cancer and lymphoma
Virus	Human papillomavirus	Cervical and anal cancers
	Hepatitis B and C virus	Liver cancer
	Human immunodeficiency virus	Kaposi's sarcoma, non-Hodgkin's lymphoma
	Human herpesvirus type 8	Kaposi's sarcoma
	Epstein–Barr virus	Lymphomas
	Human T-cell lymphotropic virus	Adult T-cell leukaemia
Parasite	Schistosomes	Bladder cancer
	Liver flukes	Cholangiocarcinoma

*Modified from ref. 15.

raise high levels of circulating anti-tumour T cells, raises a perplexing problem. Why do cancers continue to grow in the face of seemingly potent cellular anti-tumour reactions? No clear explanation for this phenomenon exists, but many hypotheses have been proposed⁴¹.

The factors limiting the therapeutic impact of anti-tumour T cells can be divided into either lymphocyte or tumour factors. Many of the T cells that are found within tumours are CD8⁺ cells. Experimental evidence in mice, as well as preliminary evidence in humans, suggests that the survival and effectiveness of CD8⁺ cells is dependent on helper factors derived from CD4⁺ cells²⁷. Thus, a successful immune reaction depends on the generation of both CD4⁺ and CD8⁺ cells, each of which are stimulated by unique and separate antigens. The general technique for cloning CD4⁺ cells described earlier will be of value in discovering antigens that can be used to stimulate CD4⁺ helper reactions¹⁴.

Although T cells can be found that react against tumour, these may be present at insufficient levels to mediate tumour destruction. The immune reaction against EBV antigens in patients with infectious mononucleosis can rise as high as 40% of all circulating CD8⁺ cells, and such large numbers of anti-tumour T cells may be required to achieve anti-tumour effects. It is also possible that the T cells that are generated do not have sufficient avidity for tumour cells, or that the T cells that are generated do not produce the appropriate cytokines or have sufficient lytic activity. To study these phenomena, efforts at cloning T lymphocytes with especially high avidity for tumour cells or unique immunological functions will aid in the understanding of the types of immune cells that are required for successful anti-tumour immune responses^{27–29}.

There are a variety of active mechanisms that may limit the effectiveness of immune stimulation. These include: active 'tolerance' of T cells resulting from the lack of expression of appropriate co-stimulatory molecules on the tumour; the active downregulation of T-cell-receptor signal transduction; the programmed cell death (apoptosis) of T cells when encountering tumour; or an active suppression by lymphocytes.

The tumour itself may be an active participant in causing immune suppression (for example, by producing local immunosuppressive factors, such as transforming growth factor- β) and there is evidence that tumours can lose expression of tumour or HLA antigens by mechanisms of immune selection. Lack of expression by the tumour of appropriate activation factors or lack of internal cellular mechanisms for apoptosis or other cell-destruction pathways may also protect the tumour cell from immune destruction.

Concluding comments

Studies of tumour immunology and immunotherapy have entered the mainstream of current studies in immunology and cancer research. The demonstration that even bulky invasive tumours can undergo complete regression under appropriate immune stimulation by IL-2 has shown that it is indeed possible to treat cancer successfully by immune manipulation. The recent discoveries of tumour antigens, and of successful means for raising anti-tumour T-cell numbers in humans by immunization, have solved some of the problems confronting the successful application of immunotherapy to the treatment of human cancer. Current studies are aimed at optimizing immunization and understanding the mechanisms used by the tumour to escape destruction. □

1. Hewitt, H. B., Blake, E. R. & Walder, A. S. A critique of the evidence for active host defence against cancer, based on personal studies of 27 murine tumours of spontaneous origin. *Br. J. Cancer* **33**, 241 (1976).
2. Woglom, W. H. Immunity to transplantable tumors. *Cancer Res.* **4**, 129 (1929).
3. Rosenberg, S. A. (ed.) *Principles and Practice of the Biologic Therapy of Cancer* (Lippincott, Philadelphia, 2000).
4. Rosenberg, S. A. *et al.* Observations on the systemic administration of autologous lymphokine-activated killer cells and recombinant interleukin-2 to patients with metastatic cancer. *N. Engl. J. Med.* **313**, 1485–1492 (1985).
5. Rosenberg, S. A., Yang, J. C., White, D. E. & Steinberg, S. M. Durability of complete responses in patients with metastatic cancer treated with high-dose interleukin-2. *Ann. Surg.* **228**, 307–319 (1998).
6. Fyfe, G. *et al.* Results of treatment of 255 patients with metastatic renal cell carcinoma who received high dose proleukin interleukin-2 therapy. *J. Clin. Oncol.* **13**, 688–696 (1995).
7. Atkins, M. B. *et al.* High-dose recombinant interleukin-2 therapy for patients with metastatic melanoma: analysis of 270 patients treated between 1985 and 1993. *J. Clin. Oncol.* **17**, 2105–2116 (1999).
8. Rosenberg, S. A. A new era for cancer immunotherapy based on the genes that encode cancer antigens. *Immunity* **10**, 281–287 (1999).
9. Boon, T., Coulie, P. G. & Van den Eynde B. Tumor antigens recognized by T cells. *Immunol. Today* **18**, 267–268 (1997).
10. Hunt, D. F. *et al.* Characterization of peptides bound to the Class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**, 1261–1263 (1992).
11. Cox, A. L. *et al.* Identification of a peptide recognized by five melanoma-specific human cytotoxic T cell lines. *Science* **264**, 716–719 (1994).
12. Kawashima, I., Hudson, S. J. & Tsai, V. The multi-epitope approach for immunotherapy for cancer: identification of several CTL epitopes from various tumor-associated antigens expressed on solid epithelial tumors. *Hum. Immunol.* **59**, 1–14 (1989).
13. Chen, Y. T., Scanlan, M. J. & Sahin, U. A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening. *Proc. Natl Acad. Sci. USA* **94**, 1914–1918 (1997).
14. Wang, R.-F., Wang, X., Atwood, A. L., Topalian, S. L. & Rosenberg, S. A. Cloning genes encoding MHC class II-restricted antigens: mutated CDC27 as a tumor antigen. *Science* **284**, 1351–1354 (1999).
15. Lowy, D. R. & Schiller, J. T. In *Cancer Principles & Practice of Oncology* 6th edn (eds DeVita, V. T., Hellman, S. & Rosenberg, S. A.) 3189–3195 (Lippincott, Philadelphia, 2001).
16. Stoler, D. L. *et al.* The onset and extend of genomic instability in sporadic colorectal tumor progression. *Proc. Natl Acad. Sci. USA* **26**, 15121–15126 (1999).
17. Landsteiner, K. & Chase, M. W. Experiments on transfer of cutaneous sensitivity to simple compounds. *Proc. Soc. Exp. Biol. Med.* **49**, 688 (1942).
18. Klein, E. & Sjogren, H. O. Humoral and cellular factors in homograft and isograft immunity. *Cancer Res.* **20**, 452 (1960).
19. Old, L. J., Boyse, E. A. & Clarke, D. A. Antigenic properties of chemically induced tumors. *Ann. NY Acad. Sci.* **101**, 80 (1962).
20. Rosenberg, S. A. *et al.* A progress report on the treatment of 157 patients with advanced cancer using lymphokine activated killer cells and interleukin-2 or high dose interleukin-2 alone. *N. Engl. J. Med.* **316**, 889–897 (1987).
21. Muul, L. M., Spiess, P. J., Director, E. P. & Rosenberg, S. A. Identification of specific cytolytic immune responses against autologous tumor in humans bearing malignant melanoma. *J. Immunol.* **138**, 989–995 (1987).
22. Itoh, K., Platsoucas, D. C. & Balch, C. M. Autologous tumor-specific cytotoxic T lymphocytes in the infiltrate of human metastatic melanomas: activation by interleukin 2 and autologous tumor cells and involvement of the T cell receptor. *J. Exp. Med.* **168**, 1419–1441 (1988).
23. Rosenberg, S. A. *et al.* Use of tumor-infiltrating lymphocytes and interleukin-2 in the immunotherapy of patients with metastatic melanoma. Preliminary report. *N. Engl. J. Med.* **319**, 1676–1680 (1988).
24. Rosenberg, S. A. *et al.* Treatment of patients with metastatic melanoma using autologous tumor-infiltrating lymphocytes and interleukin-2. *J. Natl Cancer Inst.* **86**, 1159–1166 (1994).
25. Papadopoulos, E. B. *et al.* Infusions of donor leukocytes to treat Epstein-Barr virus-associated lymphoproliferative disorders after allogeneic bone marrow transplantation. *N. Engl. J. Med.* **17**, 1185–1191 (1994).
26. Rosenberg, S. A. *et al.* Immunologic and therapeutic evaluation of a synthetic tumor associated peptide vaccine for the treatment of patients with metastatic melanoma. *Nature Med.* **4**, 321–327 (1998).
27. Walter, E. A. *et al.* Reconstitution of cellular immunity against cytomegalovirus in recipients of allogeneic bone marrow by transfer of T-cell clones from the donor. *N. Engl. J. Med.* **333**, 1038–1044 (1995).
28. Yee, C., Gilbert, M. J. & Riddell, S. R. Isolation of tyrosinase-specific CD8+ and CD4+ T cell clones from the peripheral blood of melanoma patients following in vitro stimulation with recombinant vaccinia virus. *J. Immunol.* **157**, 4079–4086 (1996).
29. Dudley, M. E., Ngo, L. T., Westwood, J., Wunderlich, J. R. & Rosenberg, S. A. T cell clones from melanoma patients immunized against an anchor-modified gp100 peptide display discordant effector phenotypes. *Cancer J. Sci. Am.* **6**, 69–77 (2000).
30. Rosenberg, S. A. Gene therapy for cancer. *J. Am. Med. Assoc.* **268**, 2416–2419 (1992).
31. Rosenberg, S. A. *et al.* Immunizing patients with metastatic melanoma using recombinant adenoviruses encoding MART-1 or gp100 melanoma antigens. *J. Natl Cancer Inst.* **90**, 1894–1900 (1998).
32. Marshall, J. L. *et al.* Phase I study in advanced cancer patients of a diversified prime-and-boost vaccination protocol using recombinant vaccinia virus and recombinant nonreplicating avipox virus to elicit anti-carcinoma antigen immune responses. *J. Clin. Oncol.* **23**, 3963–3973 (2000).
33. Eder, J. P. *et al.* A phase I trial of a recombinant vaccinia virus expressing prostate-specific antigen in advanced prostate cancer. *Clin. Cancer Res.* **5**, 1632–1638 (2000).
34. Marshall, J. L. *et al.* Phase I study in cancer patients of a replication-defective avipox recombinant vaccine that expresses human carcinoembryonic antigen. *J. Clin. Oncol.* **17**, 332–337 (1999).
35. Restifo, N. P., Ying, H., Hwang, L. & Leitner, W. W. The promise of nucleic acid vaccines. *Gene Ther.* **2**, 89–92 (2000).
36. Wang, R. *et al.* Induction of antigen-specific cytotoxic T lymphocytes in humans by a malaria DNA vaccine. *Science* **282**, 476–480 (1998).
37. Dallal, R. M., Mailliard, R. & Lotze, M. T. In *Principles and Practice of the Biologic Therapy of Cancer* 3rd edn (ed. Rosenberg, S. A.) 705–721 (Lippincott, Philadelphia, 2000).
38. Gong, J. *et al.* Fusions of human ovarian carcinoma cells with autologous or allogeneic dendritic cells induce antitumor immunity. *J. Immunol.* **3**, 1705–1711 (2000).
39. Parkhurst, M. R. *et al.* Improved induction of melanoma reactive CTL with peptides from the melanoma antigen gp100 modified at HLA-A* 0210 binding residues. *J. Immunol.* **157**, 2539–2548 (1996).
40. Rosenberg, S. A. *et al.* Impact of cytokine administration on the generation of antitumor reactivity in patients with metastatic melanoma receiving a peptide vaccine. *J. Immunol.* **163**, 1690–1695 (1999).
41. Marincola, F. M. In *Principles and Practice of the Biologic Therapy of Cancer* 3rd edn (ed. Rosenberg, S. A.) 601–617 (Lippincott, Philadelphia, 2000).
42. Kawakami, Y. *et al.* Cloning of the gene coding for a shared human melanoma antigen recognized by autologous T cells infiltrating into tumor. *Proc. Natl Acad. Sci. USA* **91**, 3515–3519 (1994).
43. Kawakami, Y. *et al.* Identification of a human melanoma antigen recognized by tumor infiltrating lymphocytes associated with in vivo tumor rejection. *Proc. Natl Acad. Sci. USA* **91**, 6458–6462 (1994).
44. Brichard, V. *et al.* The tyrosinase gene codes for an antigen recognized by autologous cytolytic T lymphocytes on HLA-A2 melanomas. *J. Exp. Med.* **178**, 489–495 (1993).
45. Wang, R. F., Robbins, P. F., Kawakami, Y., Kang, X. Q. & Rosenberg, S. A. Identification of a gene encoding a melanoma tumor antigen recognized by HLA-A31-restricted tumor-infiltrating lymphocytes. *J. Exp. Med.* **181**, 799–804 (1995).
46. Wang, R.-F., Appella, E., Kawakami, Y., Kang, X. & Rosenberg, S. A. Identification of TRP-2 as a human tumor antigen recognized by cytotoxic T lymphocytes. *J. Exp. Med.* **184**, 2207–2216 (1996).
47. Salazar-Onfray, F. *et al.* Synthetic peptides derived from the melanocyte-stimulating hormone receptor MC1R can stimulate HLA-A2-restricted cytotoxic T lymphocytes that recognize naturally processed peptides on human melanoma cells. *Cancer Res.* **57**, 4348–4355 (1997).
48. Van der Bruggen, P. *et al.* A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science* **254**, 1643–1647 (1991).
49. Visseren, M. J. *et al.* Identification of HLA-A*0201-restricted CTL epitopes encoded by the tumor-specific MAGE-2 gene product. *Int. J. Cancer* **73**, 125–130 (1997).
50. Gaugler, B. *et al.* Human gene MAGE-3 codes for an antigen recognized on a melanoma by autologous cytolytic T lymphocytes. *J. Exp. Med.* **179**, 921–930 (1994).
51. Panelli, M. C. *et al.* A tumor-infiltrating lymphocyte from a melanoma metastasis with decreased expression of melanoma differentiation antigens recognizes MAGE-12. *J. Immunol.* 4382–4392 (2000).
52. Boel, P. *et al.* BAGE: a new gene encoding an antigen recognized on human melanomas by cytolytic T lymphocytes. *Immunity* **2**, 167–175 (1995).
53. Van Den Eynde, B. *et al.* A new family of genes coding for an antigen recognized by autologous cytolytic T lymphocytes on a human melanoma. *J. Exp. Med.* **182**, 689–698 (1995).
54. Jager, E., Chen, Y. T. & Drijfhout, J. W. Simultaneous humoral and cellular immune response against cancer-testis antigen NY-ESO-1: definition of human histocompatibility leukocyte antigen (HLA)-A2-binding peptide epitopes. *J. Exp. Med.* **187**, 265–270 (1998).
55. Wang, R.-F. *et al.* A breast and melanoma-shared tumor antigenic peptides translated from different open reading frames. *J. Immunol.* **161**, 3596–3606 (1998).
56. Robbins, P. F. *et al.* A mutated B-catenin gene encodes a melanoma-specific antigen recognized by tumor-infiltrating lymphocytes. *J. Exp. Med.* **183**, 1185–1192 (1996).
57. Chiari, R. *et al.* Two antigens recognized by autologous cytolytic T lymphocytes on a melanoma result from a single point mutation in an essential housekeeping gene. *Cancer Res.* **22**, 5785–5792 (1999).
58. Wolfel, T. *et al.* A p16INK4A-insensitive CDK4 mutant targeted by cytolytic T lymphocytes in a human melanoma. *Science* **269**, 1281–1284 (1995).
59. Mandruzzato, S., Brasseur, F. & Andry, G. A CASP-8 mutation recognized by cytolytic T lymphocytes on a human head and neck carcinoma. *J. Exp. Med.* **186**, 785–793 (1997).
60. Gueguen, M., Matard, J. J. & Gaugler, B. An antigen recognized by autologous CTLs on a human bladder carcinoma. *J. Immunol.* **160**, 6188–6194 (1998).
61. Brandle, D., Brasseur, F. & Weynants, P. A mutated HLA-A2 molecule recognized by autologous cytotoxic T lymphocytes on a human renal cell carcinoma. *J. Exp. Med.* **183**, 2501–2508 (1996).
62. Butterfield, L. H. *et al.* Generation of human T-cell responses to an HLA-A2.1-restricted peptide epitope derived from alpha-fetoprotein. *Cancer Res.* **59**, 3134–3142 (1999).
63. Vonderheide, R. H., Hahn, W. C., Schultze, J. L. & Nadler, L. M. The telomerase catalytic subunit is a widely expressed tumor-associated antigen recognized by cytotoxic T lymphocytes. *Immunity* **10**, 673–679 (1999).
64. Vissers, J. L. *et al.* The renal cell carcinoma-associated antigen G250 encodes a human leukocyte antigen (HLA)-A2.1-restricted epitope recognized by cytotoxic T lymphocytes. *Cancer Res.* **59**, 5554–5559 (1999).
65. Jerome, K. R. *et al.* Cytotoxic T-lymphocytes derived from patients with breast adenocarcinoma recognize an epitope present on the protein core of a mucin molecule preferentially expressed by malignant cells. *Cancer Res.* **51**, 2908–2916 (1991).
66. Tsang, K. Y., Zaremba, S. & Nieroda, C. A. Generation of human cytotoxic T cells specific for human carcinoembryonic antigen epitopes from patients immunized with recombinant vaccinia-CEA vaccine. *J. Natl Cancer Inst.* **87**, 982–990 (1995).
67. Theobald, M. B. J., Dittmer, D., Levine, A. J. & Sherman, L. A. Targeting p53 as a general tumor antigen. *Proc. Natl Acad. Sci. USA* **92**, 11993–11997 (1995).
68. Ioannides, C. G. *et al.* T cells isolated from ovarian malignant ascites recognize a peptide derived from the HER-2/neu proto-oncogene. *Cell Immunol.* **151**, 225–234 (1993).
69. Li, K. *et al.* Tumour-specific MHC-class-II-restricted responses after in vitro sensitization to synthetic peptides corresponding to gp100 and annexin II eluted from melanoma cells. *Cancer Immunol. Immunother.* **47**, 32–38 (1998).
70. Chaux, P. *et al.* Identification of MAGE-3, epitopes presented by HLA-DR molecules to CD4+ T lymphocytes. *J. Exp. Med.* **189**, 767–777 (1999).
71. Topalian, S. L. *et al.* Human CD4+ T cells specifically recognize a shared melanoma-associated antigen encoded by the tyrosinase gene. *Proc. Natl Acad. Sci. USA* **91**, 9461–9465 (1994).
72. Zeng, G. *et al.* Identification of CD4+ T cell epitopes from NY-ESO-1 presented by HLA-DR molecules. *J. Immunol.* **165**, 1153–1159 (2000).
73. Pieper, R. *et al.* Biochemical identification of a mutated human melanoma antigen recognized by CD4+ T cells. *J. Exp. Med.* **189**, 757–766 (1999).
74. Wang, R.-F., Wang, X. & Rosenberg, S. A. Identification of a novel major histocompatibility complex class II-restricted tumor antigen resulting from a chromosomal rearrangement recognized by CD4+ T cells. *J. Exp. Med.* **189**, 1659–1667 (1999).

Haematopoietic cell transplantation as immunotherapy

Frederick R. Appelbaum

Clinical Research Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, D5-310, PO Box 19024, Seattle, Washington 98109-1024, USA

The graft-versus-tumour effect seen after allogeneic (genetically different) haematopoietic cell transplantation for human malignancies represents the clearest example of the power of the human immune system to eradicate cancer. Recent advances in our understanding of the immunobiology of stem-cell engraftment, tolerance and tumour eradication are allowing clinicians to better harness this powerful effect.

High-dose systemic chemotherapy or chemoradiotherapy followed by allogeneic haematopoietic cell transplantation (HCT) can be an effective therapy for many patients with otherwise fatal haematological malignancies. Cure rates vary, but range from as high as 80% for patients with chronic myeloid leukaemia (CML) treated during chronic phase to only 15–20% for patients with acute leukaemia refractory to conventional chemotherapy. Despite its limitations and toxicity, allogeneic transplantation is sufficiently effective that it was used to treat approximately 18,000 patients worldwide last year alone.

The initial rationale for HCT came from laboratory and clinical observations that most haematological malignancies exhibit a steep dose–response reaction to alkylating agents and radiation therapy. Because marrow toxicity is dose limiting for many of these agents, by transplanting pluripotent haematopoietic stem cells contained in bone marrow or peripheral blood, it became possible to administer far higher doses of therapy than was otherwise possible. It has since become apparent that immunocompetent cells transplanted with the stem cells, or arising from them, exert a potent graft-versus-tumour effect independent of the effects of the high-dose therapy.

Barnes *et al.* first suggested the existence of a graft-versus-tumour (GVT) effect in 1956 when they noted eradication of leukaemia in irradiated mice receiving allogeneic marrow transplants, but not syngeneic transplants from identical twins¹. The initial evidence for such an effect in humans came from studies reporting that relapse rates following allogeneic transplantation were markedly less in patients who developed graft-versus-host disease (GVHD) compare with those who did not^{2,3}. Subsequent studies revealed that relapse rates are least in patients who develop both acute and chronic GVHD, higher in those who develop no clinically evident GVHD, and higher still if T cells are depleted from the marrow graft or in recipients of twin transplants (see Fig. 1)⁴. Further verification of the GVT effect came from efforts to treat patients for post-transplant leukaemic recurrence by infusing donor lymphocytes in hopes of inducing GVHD and an associated anti-tumour effect. Somewhat surprisingly, sustained complete responses were seen in most patients with CML, and in many patients with other haematological malignancies⁵. With increased recognition of the strength of the GVT effect and the recent development of methods to better exploit it, clinical

research is beginning to focus on allogeneic HCT more as an immunotherapeutic approach, rather than solely a vehicle to deliver high-dose therapy.

Histocompatibility and transplantation

Transplantation of allogeneic haematopoietic cells is accompanied by reciprocal immunological reactions of the graft against its new host and the host against the graft. The human leukocyte antigen (HLA) system, which is the human analogue of a multigene system known as the major histocompatibility complex (MHC), is crucial in the development of these reactions⁶. Located on chromosome 6, it spans more than 4 megabases and includes more than 200 genes. For allogeneic HCT, the most influential genes are *HLA-A*, *HLA-B* and *HLA-C*, collectively referred to as class I genes, and *DRB1*, *DQB1* and *DPB1*, collectively referred to as class II genes. The class I genes are expressed on virtually all nucleated cells, whereas expression of class II genes is restricted largely to cells of the immune system. These genes are highly polymorphic; more than 125 *HLA-A*, 260 *HLA-B*, 75 *HLA-C*, 225 *HLA-DRB1* and 40 *HLA-DQB1* alleles have been described⁷.

HLA molecules are fundamental in T-cell activation, as they bind peptides and present them to T cells. HLA class I molecules preferentially present peptides to CD8⁺ T cells, whereas CD4⁺ T cells preferentially recognize peptides presented by HLA class II molecules^{8,9}. The HLA molecules themselves are termed major histocompatibility antigens and T cells confronting non-identical HLA molecules react vigorously. The peptides presented by HLA molecules may come from external sources (for example, viruses), but mostly they derive from endogenous proteins. During normal maturation of the immune system, tolerance develops to these 'self' proteins. However, in the context of organ transplantation, polymorphisms in these endogenous proteins serve as sources of minor histocompatibility antigens and form the basis of immunological non-identity between HLA-matched individuals¹⁰.

The immunological non-identity between donor and recipient has three main consequences for the use of allogeneic HCT as immunotherapy. First, after transplantation, the host may mount an immunological attack against the graft, leading to graft rejection. Because engraftment is required to exert a GVT response, methods to assure sustained engraftment are necessary. Second, immunocompetent cells in the graft can react against antigens of normal host tissues, which can result in life-threatening or even fatal GVHD. For the safe application of HCT, this reaction must

be controlled. Third, the GVT effect has been closely intertwined with the development of GVHD. If GVHD is to be controlled and the GVT effect strengthened, strategies to separate the two are required.

Engraftment and non-myeloablative transplants

Studies performed more than three decades ago using outbred species matched for major histocompatibility antigens found that very high dose chemotherapy or systemic radiotherapy administered to the recipient pretransplant was necessary to eradicate host T cells sufficiently to prevent graft rejection¹¹. Thus, until recently, most transplant 'preparative' regimens included marrow-ablative doses of therapy, not only for their anti-tumour effect, but also to ensure sustained engraftment. The intensity of these regimens limited the application of transplantation to younger, relatively healthy patients and made it difficult to distinguish the anti-tumour effects of the graft from those of the intensive preparative regimen.

With the development of more specifically immunosuppressive chemotherapeutic agents, such as fludarabine, and increased appreciation of the GVT effect, investigators have begun exploring less intensive 'non-myeloablative' preparative regimens, for example, fludarabine with moderately high dose melphalan or busulphan. Initial studies report sustained engraftment in recipients of grafts from HLA-matched siblings, diminished toxicity compared to conventional approaches, and long-term disease-free survival in a proportion of patients^{12–14}.

Although these studies focused largely on pretransplant cytotoxic therapy to enable engraftment, post-transplant treatment of the graft recipient with potent immunosuppression contributes significantly to preventing graft rejection. One set of experiments is shown in Table 1. Using the model of DLA (the canine equivalent of HLA)-identical littermates, Storb *et al.* showed that if no post-transplant immunosuppression is given, dogs require 920 cGy total body irradiation (TBI) to engraft. But if two potent immunosuppressive agents are given post-transplant, the dose of TBI required to achieve engraftment falls to 200 cGy, a dose far below that which causes bone marrow aplasia¹⁵. These observations prompted studies of conditioning regimens of very limited intensity in humans, to determine whether, as in animals, engraftment could be achieved with such low-dose therapy, and if so, what extent of tumour response would follow.

Initial clinical trials by our group involved patients who were not candidates for conventional transplants because of age or other medical problems, but who had haematological malignancies that were otherwise appropriate for transplantation. The initial treatment plan followed the animal model, and involved pretransplant treatment of patients with 200 cGy TBI followed post-transplant by the administration of mycophenolate mofetil and cyclosporine. Because occasional cases of graft rejection were seen in the first cohort of patients, low-dose fludarabine was added to the pretransplant regimen. Once fludarabine was added, graft rejection ceased to be a problem. Results in the first 109 patients have so far been reported^{16,17}. These patients (median age 55) had a variety of otherwise incurable haematological malignancies, but tolerated the transplant procedure well. Fifty-seven percent were treated entirely as outpatients, with the remaining requiring hospitalizations averaging approximately one week, as compared with an average hospitalization of over one month with conventional transplantation. The treatment-related death rate over the first 3 months was 4.5%, substantially less than the 15–20% rates seen in younger patients treated with conventional myeloablative transplant regimens. Sixty-six percent of patients who had measurable tumour before transplant achieved a complete response with this treatment. Responses were seen in virtually all categories of haematological malignancy, but were most frequent and enduring in patients with less rapidly proliferative diseases such as CML, chronic lymphocytic leukaemia and nodular lymphoma, perhaps reflecting the kinetics of the GVT response.

Non-haematological malignancies have also responded to similar low-intensity transplant approaches. Childs *et al.* reported that 10 of

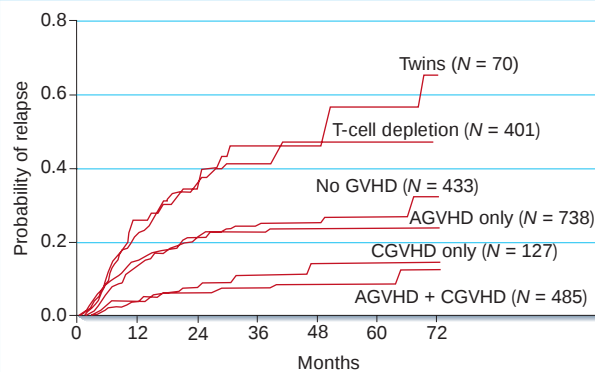


Figure 1 Relapse rates following allogeneic and syngeneic marrow transplantation. Relapse rates are least in patients who develop both acute and chronic graft-versus-host disease (AGVHD + CGVHD), higher in those who develop no clinically evident GVHD, and higher still if T cells are depleted from the marrow graft or in recipients of twin transplants⁴.

19 patients (53%) with metastatic renal-cell carcinoma exhibited disease regression, including three complete responses that have endured for periods beyond 2 years¹⁸.

These studies show that engraftment of allogeneic haematopoietic cells can be achieved with very low dose therapy and can result in pronounced anti-tumour effects. However, the procedure is also accompanied by significant GVHD in approximately 50% of individuals, and anti-tumour responses are frequently less than complete. Thus, methods both to prevent GVHD and to augment the GVT effect are required.

Induction of immunological tolerance

Because of the importance of HLA compatibility in the outcome of transplantation, most allogeneic transplants have been between HLA-matched individuals. Before 1980, this exclusively meant HLA-matched siblings, but only one in three patients have such donors available. Since that time, approximately 6.5 million normal individuals have been HLA-typed as potential unrelated volunteer marrow donors. This allows for the identification of HLA-matched unrelated donors for over 50% of patients lacking matched siblings.

GVHD results from T cells transplanted with the graft or developing from it reacting with major or minor histocompatibility antigens of the genetically different host. The development of clinically significant GVHD, although associated with a reduced risk of leukaemic relapse, leads to poorer overall survival owing to the direct effects of the disease and the consequences of the immunosuppression used to treat it¹⁹. Conventional methods to prevent GVHD have relied on a combination of the antimetabolite methotrexate given early after transplant to kill donor T cells responding vigorously to host antigens, along with cyclosporine, which blocks a calcium-dependent signal-transduction pathway distal to engagement of the T-cell receptor. Despite such prophylaxis, significant GVHD develops in 40% of patients transplanted from matched siblings and 70% of recipients of matched unrelated transplants²⁰.

The higher incidence of GVHD in recipients of unrelated transplants has variously been ascribed to unrecognized incompatibilities in major histocompatibility antigens or greater heterogeneity in minor histocompatibility antigens. Before 1998, HLA typing was largely dependent on serologic methods, which do not identify all differences. More recently, studies have been conducted in which HLA-A, -B, -C, DRB1 and DQB1 have been analysed at the allele level using automated direct sequencing. These studies detected allele-level mismatches in over 30% of serologically matched donor–recipient pairs²¹. Allele-level mismatching at class I antigens is associated with an increased incidence of graft rejection, but has no

impact on GVHD, whereas mismatching at class II is associated with increased GVHD without effect on graft rejection^{21,22}. Overall survival is markedly less in patients with multiple class I allele-level mismatches and in those with both class I and class II mismatches. Although allele-level matching should improve survival, completely matched donors will not be available for all patients, and so efforts are being made to identify those mismatches that are permissive and not associated with increased GVHD or graft rejection.

A substantial amount of work has focused on T-cell depletion (TCD) of the donor stem-cell graft as a method of preventing GVHD²³. A number of techniques exist for removing T cells, most of which use antibodies (complement mediated lysis, immunotoxins and immunomagnetic beads) or physical methods (soybean lectin agglutination, counter-flow elutriation and albumin-gradient fractionation). Clinical studies using these approaches have shown unambiguously that TCD markedly reduces the incidence and severity of GVHD. However, TCD is associated with an increased rate of severe and often fatal infections, a higher incidence of graft rejection, and an increased risk of leukaemia recurrence.

The increase in infectious complications is explained by studies showing that very few CD4⁺ and CD8⁺ T cells develop from the transplanted stem cell over the first three months after transplant and, therefore, T-cell immunity during this period is dependent on T cells transplanted with the stem cells²⁴. The increased incidence of graft failure with TCD probably reflects the loss of the contribution that donor T cells normally make in eradicating residual host immune cells surviving the transplant preparative regimen. The increased leukaemia relapse rate seen after T-cell depletion (Fig. 1) highlights the importance of the T-cell response in eradicating malignancy. The impact of TCD on graft rejection can be reduced by further intensifying the preparative regimen with additional chemotherapy and antithymocyte globulin²⁵. To lessen the impact of TCD on infections and leukaemia relapse, partial TCD, delayed re-infusion of donor lymphocytes, and post-transplant administration of low-dose interleukin (IL)-2 are all being studied^{26–28}.

As an alternative to TCD, techniques capable of inducing antigen-specific tolerance shortly after allogeneic HCT are conceptually

appealing in that they would prevent GVHD without resulting in profound post-grafting immunosuppression. One approach to the development of antigen-specific tolerance builds on the observation in murine models that exposure of antigen-activated T cells to antibodies against the invariant CD3 domain of the T-cell receptor can induce apoptosis specifically in activated cells, thereby preventing GVHD²⁹. Accordingly, a humanized non-FcR-binding anti-CD3 antibody, Hu291, has been developed and is now in clinical trials with promising early results.

A second approach for the development of antigen-specific tolerance is based on the 'two-signal model' of T-cell activation. T-cell activation requires not only stimulation of the T-cell receptor with its appropriate antigen in the context of MHC, but also a second 'co-stimulatory' signal provided by CD28 (Fig. 2). Experiments showed that stimulation of T cells with antigen plus an activating CD28 antibody *in vitro* induced IL-2 gene expression, whereas a blocking anti-CD28 antibody caused inactivation of the IL-2 gene³⁰. And transplantation using CD28-knockout mice as donors resulted in partial protection of recipients from lethal GVHD³¹. CD28 binds to two ligands, B7-1 (CD80) and B7-2 (CD86) on antigen-presenting cells (APCs)³². These ligands bind an additional T-cell antigen, CTLA4, which is expressed only after T-cell activation. A soluble CTLA4-immunoglobulin fusion protein (CTLA4-Ig) has been produced as a competitive inhibitor blocking CD28–B7 interactions. CTLA4-Ig blocks rejection of human pancreatic islet cells in mice and induces long-term, donor-specific tolerance³³. A possible shortcoming of CTLA4-Ig is that it may also block interactions of B7 and CTLA4 itself, which serves as a negative regulator of T-cell activation³⁴. Thus, an alternative approach is to block CD28–B7 interactions directly using a CD28-specific antibody. In murine models, one such antibody has been effective in preventing GVHD, but this approach has yet to be tested in humans^{35,36}.

Other studies have focused on the role of inflammatory cytokines and host APCs in the pathogenesis of GVHD. The intense preparative regimens administered before transplantation and subsequent infections induce secretion of tumour-necrosis factor (TNF)- α , IL-1 and other pro-inflammatory cytokines from APCs and other host tissues, which amplify subsequent alloimmune reactions and lead to greater GVHD³⁷. In murine models, this reaction can be substantially blocked by using specific cytokine antagonists such as IL-1-receptor antagonist or antibodies to TNF, which has led to similar trials in humans. Furthermore, although donor APCs can cross-present host antigens, they seem to do so less effectively than host APCs. Thus, pretransplant elimination of host APCs capable of presenting host antigens via the endogenous pathway can reduce GVHD in murine models³⁸. Unfortunately, no practical method exists to accomplish this in humans.

Segregating anti-tumour from anti-host reactions

The potency of the GVT effect, coupled with the direct demonstration of complete tumour responses following infusion of donor T cells for post-transplant relapse, has fuelled interest in the development of T-cell therapy to treat haematological malignancies. Clinical trials, in which T cells specific for cytomegalovirus (CMV), human immunodeficiency virus, Epstein-Barr virus and melanoma antigens were adoptively transferred, have established the safety of this approach and demonstrated that transferred T cells can persist *in vivo*, migrate to sites of antigen, and exert effector function. For example, using CMV immunity as a model, Riddell *et al.* isolated CMV-specific CD8⁺ T cells from marrow donors pre-transplant, expanded the cells *in vitro*, and administered the cells to marrow transplant recipients, which led to reconstitution of potentially protective T-cell immunity to CMV^{39,40}. Thus, many of the principles required for successful adoptive T-cell therapy have been established. The main challenge now is the identification of antigens that can be used to effectively separate anti-tumour from anti-host reactions.

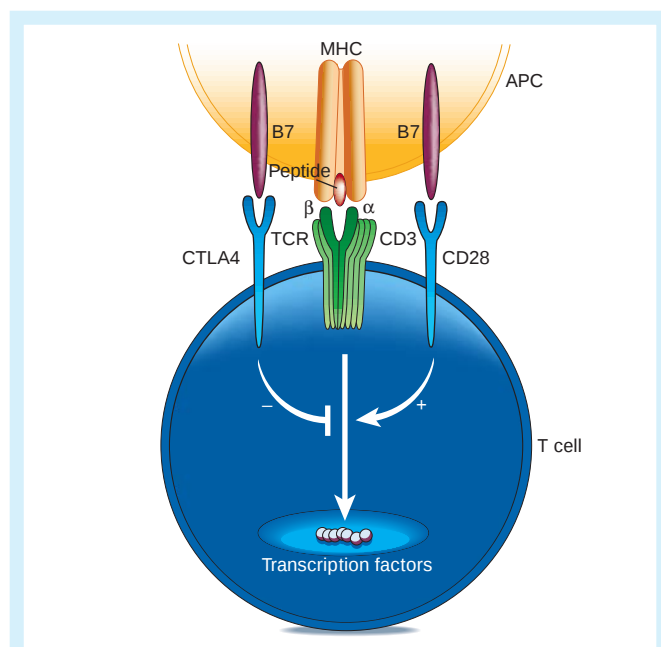


Figure 2 T-cell regulation by CD28 and CTLA4. T-cell activation requires stimulation of the T-cell receptor with its appropriate antigen presented by MHC on the antigen-presenting cell. The degree of T-cell response is regulated by secondary interactions of B7 molecules on the antigen-presenting cell with CD28, which leads to activation, or with CTLA4, which results in inhibition of response.

Two general categories of antigens capable of serving as targets for T-cell therapy are currently under study — polymorphic minor histocompatibility antigens, and antigens associated with the malignant phenotype. As noted in Fig. 1, the GVT effect can occur in the absence of GVHD, indicating that a subset of minor histocompatibility antigens (mHAgS) expressed by leukaemia cells and recognized by donor T cells are not expressed by the non-haematopoietic tissues that are targets of GVHD. Thus, one general strategy to segregate GVT from GVHD is to identify polymorphic minor histocompatibility antigens that are differentially expressed by haematopoietic and non-haematopoietic tissue. Such antigens should be able to serve as targets for post-transplant donor-derived T-cell therapy aimed at ablating all normal and malignant lymphohaematopoietic cells of the host.

In an effort to identify relevant minor histocompatibility antigens, Goulmy and Engelhard used a biochemical approach involving the elution of peptides from class I MHC molecules, separation of peptide fractions by high-performance liquid chromatography, identification of fractions that reconstitute T-cell recognition when pulsed onto target cells, and sequencing of the recognized peptides by mass spectrometry^{41,42}. Five human mHAgS have been identified by this approach, three of which are broadly expressed and thus likely to serve as targets both for GVHD and for GVT, and two that seem to be expressed selectively by haematopoietic cells and thus could serve as targets for a GVT-specific response^{43,44}. Riddell and Warren have used cell-culture techniques to isolate CD8⁺ mHAg-specific T-cell clones from allogeneic HCT recipients, and have characterized 38 previously undescribed mHAgS^{45,46}. A substantial number of these antigens seem to be restricted in their expression to haematopoietic tissues, as T-cell clones lyse recipient cells of haematopoietic origin but not recipient fibroblasts.

Acute myeloid leukaemia (AML) consists of a hierarchy of cells with different capacities for self-renewal. Transplantation of human AML into NOD/SCID mice has identified a potential leukaemia stem cell, termed the SCID leukaemia-initiating cell⁴⁷. Whether any specific mHAg identified by existing screening will be expressed by the patient's leukaemic stem cell is uncertain, but can be inferred by measuring the ability of T cells specific for such an antigen to eliminate human AML in the NOD/SCID mouse. When this was tested, outgrowth of AML was specifically prevented⁴⁸.

Clinical trials testing the safety and efficacy of T-cell clones specific for mHAgS have been initiated for patients with leukaemia in relapse after allogeneic transplantation. Because it is not possible to screen all normal tissues for expression of the targeted antigen, T-cell clones for initial infusion have been transfected with a herpes simplex thymidine-kinase suicide gene. Early results suggest that administration of such cells can be associated with achievement of complete remission without necessarily augmenting GVHD (ref. 44, and S. R. Riddell, personal communication).

Although polymorphic minor histocompatibility antigens are rational targets for separating GVT from GVHD, they will always require the context of allogeneic HCT for their use. An alternative approach is to identify antigens associated with the malignant phenotype. Such antigens could serve as targets for T-cell therapy, not only following allogeneic transplantation, but also in other settings. Candidate antigens can be classified as: (1) mutational, such as Bcr/Abl, which should be tumour specific; (2) viral, such as human papillomavirus in cervical cancer, which should also be tumour specific; (3) tissue specific, such as prostate-specific antigen in prostate cancer, where destruction of normal tissue is permissible; (4) germ-cell antigens, such as the melanoma-associated antigen family, normally expressed in adults only in the testes; and (5) overexpressed self-proteins, which may be recognized based on increased levels of presentation. The immunogenicity of such antigens is influenced by many factors, including level of protein expression, the peptide sequences resulting from intracellular processing, and the ability of these peptides to be presented by class I molecules. Host responses to such antigens may range from non-responsiveness

Table 1 Marrow grafts from DLA-identical canine littermates after TBI

TBI dose (cGy)	Postgrafting immunosuppression	No. engrafted/No. studied
920	—	20/21
450	—	16/39
450	CSP	7/7
200	CSP	0/4
200	MMF/CSP	11/12
100	MMF/CSP	0/6

Abbreviations: TBI, total body irradiation; CSP, cyclosporine; MMF, mycophenolate mofetil.

resulting from prior tolerance, to production of T cells of sufficient avidity to selectively destroy targeted cells.

Mutational antigens considered as targets for T-cell therapy include Bcr/Abl in CML and PML/RAR α in acute promyelocytic leukaemia. However, reproducible generation of CD8⁺ T cells that recognize leukaemia progenitor cells bearing these antigens has not yet been possible^{49,50}. Normal proteins overexpressed in leukaemia progenitors may provide alternative targets. Ideally, such antigens should be expressed at substantially higher levels in leukaemia than in normal cells, expressed by all leukaemia cells and be efficiently processed and presented by diverse HLA class I molecules. Two proteins that fit this description, proteinase 3 (PR3) and Wilms' tumour-suppressor (WT1), have already been shown to elicit CD8⁺ responses.

PR3 is a neutral serine proteinase with expression largely restricted to the promyelocytic stage of myeloid differentiation^{51,52}. Although PR3 is not detected in normal haematopoietic stem cells, it is expressed in freshly isolated leukaemia progenitors, particularly from CML patients. CD8⁺ T cells specific for PR3 have been generated by stimulation of cells with a peptide predicted to bind to the A2.1 class I allele^{53,54}. Such cells selectively lyse leukaemic blasts but not normal bone marrow cells. CD8⁺ cytotoxic T lymphocytes (CTLs) specific for this epitope were not detected in the peripheral blood of normal individuals or in untreated CML patients, but they could be found in the blood of CML patients who had been treated successfully with allogeneic transplantation or with α -interferon, suggesting a role for these cells in the anti-tumour response⁵⁵.

WT1, a zinc-finger transcription factor, was initially described as a tumour-suppressor gene in childhood Wilms' tumour. WT1 is abundantly overexpressed in most human leukaemia cells, including AML, CML and acute lymphocytic leukaemia, with higher levels associated with a worse prognosis^{56,57}. Leukaemia cells express from 10- to >100-fold more WT1 protein than normal CD34⁺ cells. Studies indicate that T cells can distinguish this difference in protein expression, as CD8⁺ CTLs generated against WT1 lyse leukaemic CD34⁺ but not normal CD34⁺ cells, and inhibit growth of leukaemic but not normal myeloid colonies⁵⁸. Thus, like PR3, WT1 might serve as a useful target for adoptive T-cell therapy. Development of gene microarrays, which enable the expression of thousands of genes simultaneously, should facilitate the identification of additional proteins that are overexpressed in leukaemia progenitors.

Success in the application of T-cell therapy will have important implications for alternative strategies, particularly vaccine development. Murine studies have, in fact, shown enhanced immune responses to a vaccine based on granulocyte-macrophage colony-stimulating factor administered in the early post-transplant period, compared to the non-transplant setting⁵⁹. Direct T-cell transfer should greatly aid vaccine development by defining the therapeutic efficacy of targeting any specific antigen and any possible toxicity to normal tissues.

Summary

The ability to achieve complete haematopoietic engraftment without intensive therapy will have a profound effect on the practise of allogeneic HCT. Rather than treating patients with high-dose preparative regimens to both eradicate the malignancy and prevent graft rejection, efforts to capture the benefits of high-dose therapy can

focus on developing treatments specifically targeted to tumour eradication and combining these with specific immunosuppression to ensure engraftment. For example, studies are now underway combining high-dose monoclonal antibody-targeted radiotherapy with non-myeloablative transplant regimens⁶⁰. Methods to induce antigen-specific tolerance following transplantation promise to reduce GVHD without producing severe prolonged immunodeficiency. Finally, and perhaps most important, strategies now exist to segregate GVT from GVHD. The identification of a modest number of polymorphic minor histocompatibility antigens with expression limited to the lymphohaematopoietic system should allow augmentation of the GVT response in most patients transplanted for haematological malignancies, either by adoptive transfer of T cells, or perhaps, vaccination of the donor before transplant. If the adoptive transfer of T cells specific for overexpressed tumour antigens proves safe and effective, as preclinical experiments predict, this will encourage study of adoptive T-cell transfer in both transplant and non-transplant settings, and pave the way for vaccine trials. Furthermore, with the rapid application of gene microarray analyses, additional new candidate antigens will probably become apparent for multiple tumour types. □

1. Barnes, D. W. H., Corp, M. J., Loutit, J. F. & Neal, E. E. Treatment of murine leukaemia with x-rays and homologous bone marrow. Preliminary communication. *Br. Med. J.* **2**, 626–627 (1956).
2. Weiden, P. L. et al. Antileukemic effect of graft-versus-host disease in human recipients of allogeneic-marrow grafts. *N. Engl. J. Med.* **300**, 1068–1073 (1979).
3. Weiden, P. L., Sullivan, K. M., Flournoy, N., Storb, R., Thomas, E. D. & the Seattle Marrow Transplant Team. Antileukemic effect of chronic graft-versus-host disease. Contribution to improved survival after allogeneic marrow transplantation. *N. Engl. J. Med.* **304**, 1529–1533 (1981).
4. Horowitz, M. M. et al. Graft-versus-leukemia reactions after bone marrow transplantation. *Blood* **75**, 555–562 (1990).
5. Kolb, H. J. et al. Graft-versus-leukemia effect of donor lymphocyte transfusions in marrow grafted patients. European Group for Blood and Marrow Transplantation Working Party Chronic Leukemia. *Blood* **86**, 2041–2050 (1995).
6. Klein, J. & Sato, A. The HLA system. *N. Engl. J. Med.* **343**, 702–709, 782–786 (2000).
7. Bodmer, J. G. et al. Nomenclature for factors of the HLA system, 1998. *Tissue Antigens* **53**, 407–446 (1999).
8. Doyle, C. & Strominger, J. L. Interaction between CD4 and class II MHC molecules mediates cell adhesion. *Nature* **330**, 256–259 (1987).
9. Norment, A. M., Salter, R. D., Parham, P., Engelhard, V. H. & Littman, D. R. Cell–cell adhesion mediated by CD8 and MHC class I molecules. *Nature* **336**, 79–81 (1988).
10. Parham, P. Deconstructing the MHC. *Nature* **360**, 300–301 (1992).
11. Storb, R. et al. Marrow grafts between DL-A-matched canine littermates. *Transplantation* **15**, 92–100 (1973).
12. Khouri, I. F. et al. Transplant-lite: induction of graft-versus-malignancy using fludarabine-based nonablative chemotherapy and allogeneic blood progenitor-cell transplantation as treatment for lymphoid malignancies. *J. Clin. Oncol.* **16**, 2817–2824 (1998).
13. Slavin, S., Nagler, A., Aker, M., Shapira, M. & Or, R. in *New Frontiers in Cancer Therapy: Non-myeloablative Stem Cell Transplantation (NST)* (eds Giral, S. & Slavin, S.) 83–91 (Darwin Scientific Publishing, Oxford, 2000).
14. Giral, S. et al. Melphalan and purine analog-containing preparative regimens: reduced-intensity conditioning for patients with hematologic malignancies undergoing allogeneic progenitor cell transplantation. *Blood* **97**, 631–637 (2001).
15. Storb, R., Yu, C. & McSweeney, P. in *Hematopoietic Cell Transplantation*, 2nd edn (eds Thomas, E. D., Blume, K. G. & Forman, S. J.) 287–295 (Blackwell Science, Boston, 1999).
16. McSweeney, P. A. et al. Hematopoietic cell transplantation in older patients with hematologic malignancies: replacing high-dose cytotoxic therapy with graft-versus-tumor effects. *Blood* (in the press).
17. Sandmaier, B. M. et al. Nonmyeloablative conditioning for HLA-identical related allografts for hematologic malignancies. *Blood* **96**, 479a (abstract 2062) (2000).
18. Childs, R. et al. Regression of metastatic renal-cell carcinoma after nonmyeloablative allogeneic peripheral-blood stem-cell transplantation. *N. Engl. J. Med.* **343**, 750–758 (2000).
19. Nash, R. A. et al. Acute graft-versus-host disease: analysis of risk factors after allogeneic marrow transplantation and prophylaxis with cyclosporine and methotrexate. *Blood* **80**, 1838–1845 (1992).
20. Sullivan, K. M. in *Bone Marrow Transplantation* (eds Forman, S. J., Blume, K. G. & Thomas, E. D.) 339–362 (Blackwell Scientific Publications, Boston, 1994).
21. Petersdorf, E. W. et al. Optimizing outcome after unrelated marrow transplantation by comprehensive matching of HLA class I and II alleles in the donor and recipient. *Blood* **92**, 3515–3520 (1998).
22. Petersdorf, E. W. et al. Association of HLA-C disparity with graft failure after marrow transplantation from unrelated donors. *Blood* **89**, 1818–1823 (1997).
23. Kernan, N. A. in *Bone Marrow Transplantation* (eds Forman, S. J., Blume, K. G. & Thomas, E. D.) 124–135 (Blackwell Scientific Publications, Boston, 1994).
24. Storek, J. & Storb, R. T-cell reconstitution after stem-cell transplantation—by which organ? *Lancet* **355**, 1843–1844 (2000).
25. Papadopoulos, E. B. et al. T-cell-depleted allogeneic bone marrow transplantation as postremission therapy for acute myelogenous leukemia: freedom from relapse in the absence of graft-versus-host disease. *Blood* **91**, 1083–1090 (1998).
26. Soiffer, R. J., Murray, C., Gonin, R. & Ritz, J. Effect of low-dose interleukin-2 on disease relapse after T-cell depleted allogeneic bone marrow transplantation. *Blood* **84**, 964–971 (1994).
27. Champlin, R. et al. Selective depletion of CD8+ T lymphocytes for prevention of graft-versus-host disease after allogeneic bone marrow transplantation. *Blood* **76**, 418–423 (1990).
28. Dazzi, F. et al. Comparison of single-dose and escalating-dose regimens of donor lymphocyte infusion for relapse after allografting for chronic myeloid leukemia. *Blood* **95**, 67–71 (2000).
29. Anasetti, C., Bidwell, B. J., Martin, P. J. & Yu, X. Z. Anti-CD3 F(AB')₂ prevents graft-versus-host disease by selectively depleting donor T cells activated by recipient alloantigens. *Blood* **96**, 582a (abstract 2499) (2000).
30. Harding, F. A., McArthur, J. G., Gross, J. A., Raulet, D. H. & Allison, J. P. CD28-mediated signalling co-stimulates murine T cells and prevents induction of anergy in T-cell clones. *Nature* **356**, 607–609 (1992).
31. Yu, X.-Z., Martin, P. J. & Anasetti, C. Role of CD28 in acute graft-versus-host disease. *Blood* **92**, 2963–2970 (1998).
32. Lenschow, D. J., Walunas, T. L. & Bluestone, J. A. CD28/B7 system of T cell costimulation. *Annu. Rev. Immunol.* **14**, 233–258 (1996).
33. Lenschow, D. J. et al. Long-term survival of xenogeneic pancreatic islet grafts induced by CTLA4lg. *Science* **257**, 789–792 (1992).
34. Perez, V. L. et al. Induction of peripheral T cell tolerance in vivo requires CTLA-4 engagement. *Immunity* **6**, 411–417 (1997).
35. Yu, X. Z., Bidwell, S. J., Martin, P. J. & Anasetti, C. CD28-specific antibody prevents graft-versus-host disease in mice. *J. Immunol.* **164**, 4564–4568 (2000).
36. Greenberg, P. D. & Riddell, S. R. Deficient cellular immunity-finding and fixing the defects. *Science* **285**, 546–551 (1999).
37. Krenger, W., Hill, G. R. & Ferrara, J. L. Cytokine cascades in acute graft-versus-host disease. *Transplantation* **64**, 553–558 (1997).
38. Shlomchik, W. D. et al. Prevention of graft versus host disease by inactivation of host antigen-presenting cells. *Science* **285**, 412–415 (1999).
39. Riddell, S. R. et al. Restoration of viral immunity in immunodeficient humans by the adoptive transfer of T cell clones. *Science* **257**, 238–241 (1992).
40. Walter, E. A. et al. Reconstitution of cellular immunity against cytomegalovirus in recipients of allogeneic bone marrow by transfer of T-cell clones from the donor. *N. Engl. J. Med.* **333**, 1038–1044 (1995).
41. Hunt, D. F. et al. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**, 1261–1263 (1992).
42. Goulmy, E. Human minor histocompatibility antigens: new concepts for marrow transplantation and adoptive immunotherapy. *Immunol. Rev.* **157**, 125–140 (1997).
43. den Haan, J. M. et al. The minor histocompatibility antigen HA-1: a diallelic gene with a single amino acid polymorphism. *Science* **279**, 1054–1057 (1998).
44. Mutis, T. et al. Feasibility of immunotherapy of relapsed leukemia with ex vivo-generated cytotoxic T lymphocytes specific for hematopoietic system-restricted minor histocompatibility antigens. *Blood* **93**, 2336–2341 (1999).
45. Warren, E. H., Greenberg, P. D. & Riddell, S. R. Cytotoxic T-lymphocyte-defined human minor histocompatibility antigens with a restricted tissue distribution. *Blood* **91**, 2197–2207 (1998).
46. Warren, E. H., Gavin, M., Greenberg, P. D. & Riddell, S. R. Minor histocompatibility antigens as targets for T-cell therapy after bone marrow transplantation. *Curr. Opin. Hematol.* **5**, 429–433 (1998).
47. Lapidot, T. et al. A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
48. Bonnet, D., Warren, E. H., Greenberg, P. D., Dick, J. E. & Riddell, S. R. CD8⁺ minor histocompatibility antigen-specific cytotoxic T lymphocyte clones eliminate human acute myeloid leukemia stem cells. *Proc. Natl Acad. Sci. USA* **96**, 8639–8644 (1999).
49. Dermime, S. et al. Lack of T-cell-mediated recognition of the fusion region of the pml/RAR-alpha hybrid protein by lymphocytes of acute promyelocytic leukemia patients. *Clin. Cancer Res.* **2**, 593–600 (1996).
50. Chen, W., Qin, H., Reese, V. A. & Cheever, M. A. CTLs specific for bcr-abl joining region segment peptides fail to lyse leukemia cells expressing p210 bcr-abl protein. *J. Immunother.* **21**, 257–268 (1998).
51. Sturrock, A., Franklin, K. F. & Hoidal, J. R. Human proteinase-3 expression is regulated by PU.1 in conjunction with a cytidine-rich element. *J. Biol. Chem.* **271**, 32392–32402 (1996).
52. Sturrock, A. B. et al. Structure, chromosomal assignment, and expression of the gene for proteinase-3. The Wegener's granulomatosis autoantigen. *J. Biol. Chem.* **267**, 21193–21199 (1992).
53. Molldrem, J. J. et al. Cytotoxic T lymphocytes specific for a nonpolymorphic proteinase 3 peptide preferentially inhibit chronic myeloid leukemia colony-forming units. *Blood* **90**, 2529–2534 (1997).
54. Molldrem, J. et al. Targeted T-cell therapy for human leukemia: cytotoxic T lymphocytes specific for a peptide derived from proteinase 3 preferentially lyse human myeloid leukemia cells. *Blood* **88**, 2450–2457 (1996).
55. Molldrem, J. J. et al. Evidence that specific T lymphocytes may participate in the elimination of chronic myelogenous leukemia. *Nature Med.* **6**, 1018–1023 (2000).
56. Inoue, K. et al. Aberrant overexpression of the Wilms tumor gene (WT1) in human leukemia. *Blood* **89**, 1405–1412 (1997).
57. Bergmann, L. et al. High levels of Wilms' tumor gene (wt1) mRNA in acute myeloid leukemias are associated with a worse long-term outcome. *Blood* **90**, 1217–1225 (1997).
58. Gao, L. et al. Selective elimination of leukemic CD34⁺ progenitor cells by cytotoxic T lymphocytes specific for WT1. *Blood* **95**, 2198–2203 (2000).
59. Borrello, I. et al. Sustaining the graft-versus-tumor effect through posttransplant immunization with granulocyte-macrophage colony-stimulating factor (GM-CSF)-producing tumor vaccines. *Blood* **95**, 3011–3019 (2000).
60. Appelbaum, F. R. Antibody-targeted therapy for myeloid leukemia. *Semin. Hematol.* **36**, 2–8 (1999).

Acknowledgements

The author thanks P. Greenberg, R. Storb, J. Hansen, E. Petersdorf, S. Riddell and C. Anasetti for the development of many of the concepts reviewed in this manuscript.

Cancer epidemiology in the last century and the next decade

Julian Peto

Institute of Cancer Research, Sutton, Surrey SM2 5NG, and London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

By the early 1980s, epidemiologists had identified many important causes of cancer. They had also proposed the 'multi-stage' model of cancer, although none of the hypothesized events in human carcinogenesis had then been identified. The remarkable advances in cell and molecular biology over the past two decades have transformed the scope and methods of cancer epidemiology. There have been a few new discoveries based purely on traditional methods, and many long-suspected minor risks have been estimated more precisely. But modern epidemiological studies often depend on genetic, biochemical or viral assays that had not been developed 20 years ago.

Many types of cancer vary in incidence by more than an order of magnitude between different populations, and every type is rare in some part of the world¹. The convergence towards local cancer rates seen among immigrants (Fig. 1) excludes a genetic explanation of these differences. By the 1960s, cancer epidemiologists had therefore concluded that most cancers are in principle preventable and many could be avoided by a suitable choice of lifestyle and environment². Many specific causes of cancer are now known, the most important being smoking, obesity and a few oncogenic viruses, but a large proportion of global variation for common cancers such as breast, prostate, colon and rectum remains unexplained.

Environmental and lifestyle causes of cancer

Carcinogenic effects of tobacco

The most important discovery in the history of cancer epidemiology is the carcinogenic effect of tobacco. Lung cancer incidence increases rapidly among continuing smokers³, so the risk is greatest in those who begin to smoke when young and continue throughout life. The large increase in male cigarette smoking in Britain during and after the First World War therefore caused an unprecedented epidemic among men born around 1900, and by 1955 the rate in British men aged under 55 was the highest in the world⁴. Over the five decades since British⁵ and American⁶ epidemiologists reported that 'cigarette smoking is a factor, and an important factor, in the production of carcinoma of the lung'⁵, there has been a marked reduction in tar levels of British cigarettes⁷ and in smoking among British men⁸. As a result, their lung cancer rate below age 55 has fallen by more than two-thirds since 1955; it is now among the lowest in the developed world and is still declining⁸. Similar changes occurred 20 years later in America, where cigarette smoking increased rapidly during the Second World War¹. Women in most Western countries began smoking later than men and fewer have stopped, so their lung cancer rates are either still increasing or falling less rapidly⁴. Male lung cancer rates are still increasing in most developing countries and in Eastern Europe, where consumption of cigarettes remains high and in some areas is still increasing⁴.

For many years the carcinogenic effects of tobacco were thought to be restricted largely to the lung, pancreas, bladder and kidney, and (synergistically with alcohol) the larynx,

mouth, pharynx (except nasopharynx) and oesophagus¹. More recent evidence indicates that several other types of cancer, of which the most important worldwide are stomach, liver and (probably) cervix, are also increased by smoking^{9–11}. The relative importance of different smoking-related diseases varies widely between populations, as smoking usually multiplies the background rate due to other factors. In China, where liver cancer is common, smoking causes more premature deaths from liver cancer than from heart disease (ref. 11 and Fig. 2). The overall proportion of male cancer deaths caused by smoking in China in 1990 was 22% and rising¹¹, whereas that in Britain fell from 44% in 1990 to about 36% by 2000^{4,12}.

The effect of diet and overweight

Dietary epidemiology is notoriously complex owing to the variety of foods and their many constituents and to intercorrelations and temporal changes in their patterns of use. Cancer risks in old age may also depend as much on diet in early life as on current habits¹³. Apart from drinking alcohol, consumption of various foods contaminated with aflatoxin¹⁴, and a few local customs (such as feeding Chinese-style salted fish to infants, which causes nasopharyngeal cancer¹⁴), no single dietary factor shows a strong and consistent enough effect to establish it unequivocally as an important carcinogen or anti-carcinogen¹⁵. Extensive research during the past two decades has shown that rates for various cancers correlate fairly consistently with certain aspects of diet, but opinions still differ on the strength of the evidence. Doll and Peto¹⁰ conclude that about a third of British cancer deaths might eventually prove to be avoidable by dietary change but only those due to obesity are definitely avoidable. In contrast, an American expert panel¹⁵ recently concluded that about a third of cancers worldwide would probably be prevented by adoption of their quantitative recommendations on daily consumption of various foods. Another British report made qualitatively similar recommendations but stated that current evidence is insufficient to determine optimal quantities¹³.

There is now a consensus that cancer is commoner in those who are overweight¹⁶. The evidence on weight is strongest for post-menopausal breast cancer and cancers of the endometrium, gall-bladder and kidney, but several other sites contribute to the overall cancer risk^{16,17}. About 5% (3% in men, 6% in women) of all incident cancers in the European Union might be prevented if no-one's body-mass index (BMI; weight divided by the square of height) exceeded

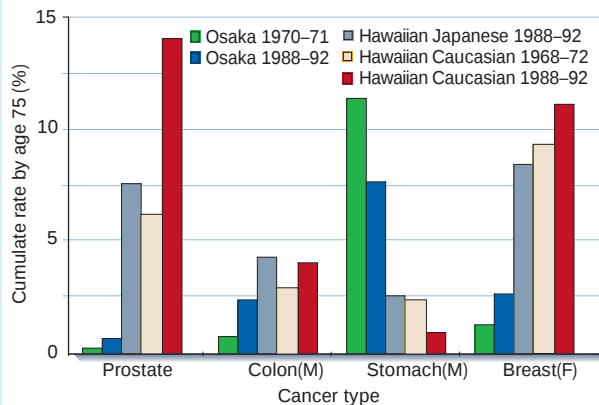


Figure 1 Cancer rates in migrants become similar to those in the local population. Cancer rates in 1990 among Japanese migrants to Hawaii, and around 1970 and 1990 in Japan (Osaka) and in Hawaiian Caucasians. Local rates for prostate, colon and breast cancer increased over time (due partly to increased completeness of diagnosis and registration, particularly for prostate cancer in Hawaiian Caucasians) and stomach cancer decreased; but the effects of migration were larger.

25 kg m⁻² (ref. 17). Exclusion of smoking-related cancers would increase this estimate to about 7%. A large prospective cohort of non-smokers in America, where obesity is more prevalent, provides the strongest evidence on BMI and cancer mortality¹⁸. The authors did not calculate an attributable fraction, but their data suggest that about 10% of all cancer deaths among American non-smokers (7% in men and 12% in women) are caused by overweight¹⁸. It is, however, not clear how much the risk can be reduced by weight reduction in those who are already overweight. Mortality from non-malignant diseases is increased in those who are either too thin or too fat¹⁸.

Radical changes in national dietary habits would not be easy to achieve even if there were a consensus on which foods are relevant. Dietary supplements such as vitamins or other micronutrients seem an attractive alternative, but they may not have the same effects as the foods that contain them, and some may even be harmful¹³. The only reliable way to assess their effectiveness is in large randomized trials that continue for many years. For example, there is substantial epidemiological evidence that consumption of foods containing beta-carotene correlates with reduced risk of lung cancer^{13,15}, but 12 years' treatment in a large randomized trial showed no benefit¹⁹, and in two shorter trials the lung cancer risk was higher in those who received beta-carotene supplements^{20,21}. Aspirin and folate supplements probably reduce colorectal cancer incidence but may take a decade or more to do so²². The American panel¹⁵ concluded that various cancers were likely to be reduced by foods containing adequate amounts of carotenoids, vitamins C and E and selenium, but neither they nor the British panel¹³ recommended that these micronutrients should be taken as supplements.

Reproductive and hormonal factors

The effects of reproductive factors on breast and ovarian cancer have long been assumed to reflect underlying hormonal processes¹, and this is confirmed by the effects of both endogenous^{23,24} and exogenous²⁵ hormones. Breast cancer incidence is transiently increased by pregnancy and while oestrogens are administered as oral contraceptives or hormone replacement therapy, and is permanently lowered by late menarche, early menopause, early first childbirth and high parity²⁵. Endometrial cancer incidence is also increased by hormone replacement therapy²⁵. Ovarian cancer incidence declines with increasing parity²⁶, and both endometrial and ovarian cancers are less common in oral contraceptive users²⁵.

The Western diet is associated both with earlier age at menarche and with post-menopausal obesity, which increases endogenous oestrogen

production and hence breast cancer risk¹⁵. Breast cancer incidence is much higher in most Western countries than in many developing countries, and this is partly (and perhaps largely) accounted for by these dietary effects combined with later first childbirth, lower parity and shorter breastfeeding^{27,28}. The development of cancers of the testis and prostate may also depend on hormonal effects, but apart from the increased risk in an undescended testis, no behavioural or reproductive correlate is strongly predictive of these diseases²⁹.

Viruses, bacteria and parasites

The most important discoveries of the past two decades in cancer epidemiology relate to the carcinogenic effects of infectious pathogens that had not been characterized 20 years ago. *Helicobacter pylori*, a chronic gastric bacterial infection that can cause gastric ulcers, is a major factor in the development of stomach cancer³⁰. More than 100 human papillomaviruses (HPVs) have been sequenced, and DNA from a phylogenetic subgroup of sexually transmitted HPVs that includes HPV16, HPV18 and HPV45 is detectable in virtually all cervical cancers worldwide³¹. These and other HPVs are also found in other anogenital cancers and may also cause cancers of other sites (head and neck, oesophagus and skin)³². The contribution of hepatitis-B virus (HBV) to liver cancer in high-incidence regions has long been recognized³³, although the synergistic effect of smoking is a more recent discovery¹¹. The hepatitis-C virus (HCV) is similarly carcinogenic³³. About one-fifth of all human cancers worldwide arise in the stomach (9%), liver (6%) or cervix (5%), and most of these would be prevented if these infections could be eradicated³⁴.

Other pathogens that cause a substantial cancer risk in certain populations include Epstein-Barr virus (EBV; associated with various B-cell malignancies and nasopharyngeal cancer), malaria (the major cofactor with EBV for Burkitt's lymphoma in Africa), human T-cell lymphotropic virus type 1 (some T-cell leukaemias and lymphomas), HIV (non-Hodgkin's lymphoma), human herpesvirus 8 (Kaposi's sarcoma, with HIV), schistosomiasis (bladder and colon cancer) and liver flukes (cholangiosarcoma)^{30,35,36}. There is also strong epidemiological evidence for an infective aetiology in childhood leukaemia³⁷, but no specific pathogen has been implicated. The incidence of several virally induced cancers is further increased by specific cofactors such as dietary aflatoxin (liver), salted fish (nasopharynx) and smoking (liver

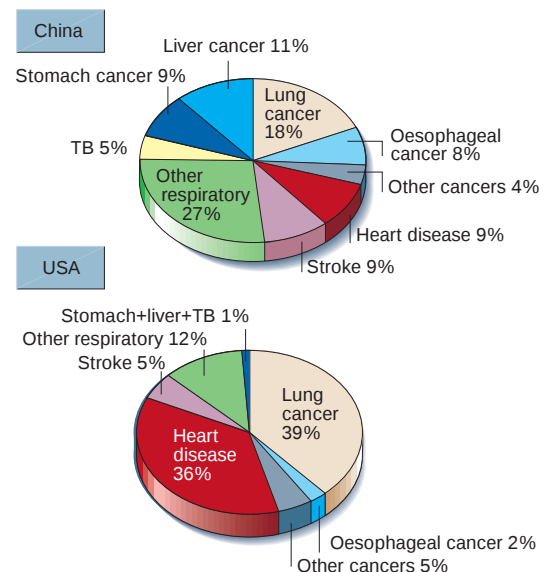


Figure 2 Smoking kills different populations in different ways. Deaths below age 70 in 1990 caused by smoking in China¹¹ and the United States⁴. 'Other cancers' are mouth, pharynx, larynx, bladder and pancreas.

and cervix), and if SV-40 is an important cause of mesothelioma, asbestos must also be classed as a cofactor³⁸.

Therapeutic immunosuppression causes a marked increase in the incidence of non-melanoma skin cancer and some virally induced cancers³⁹. The discovery that many other epithelial cancers, notably lung, colon, rectum, bladder and prostate (but not breast), are also increased by immunosuppression (Table 1 and ref. 40) suggests that unidentified viruses may be important in these cancers as well. The alternative is the long-standing but equally speculative theory that many non-viral cancers are normally kept in check by immunosurveillance (see article in this issue by Rosenberg, pages 380–384).

Occupational and environmental carcinogens

About a dozen specific occupational exposures and several complex mixtures, particularly the combustion products of coal, have caused high risks of certain cancers (predominantly lung cancer) in heavily exposed workers¹⁰. Exposure levels for many industrial hazards have been progressively reduced in many Western countries since the 1930s, and by the late 1970s it was assumed, probably correctly, that the occupational exposure levels then current would contribute a very small proportion of future cancer incidence¹. But uncontrolled asbestos use had been widespread in the European construction industry from the 1950s to the mid-1970s, when public concern led to a rapid reduction. The resulting epidemic of mesothelioma in building and other workers born after 1940 did not become apparent until the 1990s owing to the long latency of the disease. Incidence rates are still rising, and asbestos exposure prior to 1980 may eventually cause 250,000 mesotheliomas and a similar number of lung cancers in Western Europe⁴¹.

This tragic episode was largely avoidable, as the carcinogenic effects of asbestos were known by 1960^{42,43}, but it illustrates the major weakness of epidemiology as an early warning system. The increase in cancer incidence caused by increased exposure to a carcinogen might not be detectable for several decades, and laboratory testing must remain the first line of defence against potentially dangerous new agents, particularly those affecting endocrine or paracrine signalling that could be biologically active at very low levels. The unexplained increase in testicular cancer in many Western countries could be due to such compounds, although a dietary or viral explanation would also be plausible. The possibility of germ-cell damage is of particular concern, as environmental mutagens must cause some heritable changes, but the effect is so small that no known or suspected mutagen, including ionizing radiation, has measurably increased the frequency of germ-line mutation in humans⁴⁴. Epidemiological studies of markers such as DNA adducts in the lung or chromosomal aberrations in lymphocytes might also provide early warning of a potential hazard. But such direct or indirect measures of mutagenic or transforming potency have never detected an important carcinogen and even today cannot provide quantitative estimates of risk.

Epidemiological data on human cancer rates still provide the only reliable evidence that the cancer risks caused by long-established activities such as working in an oil refinery or living near a high-voltage power line are not large. Apart from skin cancers due to sunlight, the only substantial and widespread cancer risk known to be caused by an avoidable environmental factor in developed countries is the further increase in lung cancer among smokers caused by indoor radon escaping from the ground or from building materials, although both indoor and outdoor air pollution from fossil fuels may also contribute to the risk in smokers⁴⁵. The risk to non-smokers is relatively trivial in developed countries, but burning fossil fuels indoors without adequate ventilation certainly contributes to the high lung cancer rates even in non-smokers seen in parts of China¹¹.

Genetic epidemiology of cancer

Polymorphisms in candidate genes

There have been many studies comparing the prevalence in cancer patients and unaffected controls of common polymorphisms in genes involved in the metabolism of external or endogenous mutagens or in

the production or processing of sex hormones or their analogues. A few polymorphisms in such genes seem to alter the risk substantially, such as the *N*-acetyltransferase 2 (NAT2) slow acetylator phenotype, which increases the risk of bladder cancer⁴⁶, particularly in workers heavily exposed to certain aromatic amines⁴⁷. (Fast acetylators are not immune, however, as in one factory all 19 men who distilled β -naphthylamine developed bladder cancer⁴⁸.) But systematic meta-analysis reveals little or no effect for most such polymorphisms, and the pooled data for the minority that are statistically significant usually suggest odds ratios of less than two, and often much less^{46,49–51}. Thus, for example, early reports suggested a more than doubled lung cancer risk associated with glutathione *S*-transferase μ 1 (GSTM1) deficiency, but the pooled results of subsequent genotyping studies give an odds ratio of only 1.14 (95% confidence interval, 1.03–1.25)⁵⁰.

Polymorphisms in oncogenes or tumour-suppressor genes may also confer a moderately increased cancer risk. An example is the I1307K single nucleotide polymorphism (SNP) in the *APC* gene, which is carried by about 1 in 20 Ashkenazi Jews and almost doubles their colon cancer risk⁵². To estimate the individual effects of rare polymorphisms will require very large studies, but their average effect can be observed. The increased cancer risk associated with rare alleles of the *HRAS1*-associated minisatellite was among the first such associations reported. Such alleles, which are carried by about 5% of the population, increase the risk of several common cancers by a factor of 1.5 to 2 (ref. 53).

There have been various reports of statistically significant gene–environment interactions, such as a much larger lung cancer risk due to passive smoking in women who were GSTM1-deficient⁵⁴, or an increased breast cancer risk due to smoking in post-menopausal women that was confined to NAT2 slow acetylators⁵⁵. In these examples, however, the estimates of the risk in susceptibles (although not their lower confidence limits) were inconsistent with the much lower overall effect of passive smoking on lung cancer⁵⁶ or of smoking on breast cancer (which is nil) in larger studies⁵⁷. Many apparently significant gene–gene or gene–exposure interactions will arise by chance, but some will be real. The interaction between methylenetetrahydrofolate reductase and plasma folate in colorectal cancer is a plausible candidate⁵⁸. The effects of such polymorphisms in combination with each other and with environmental risk factors

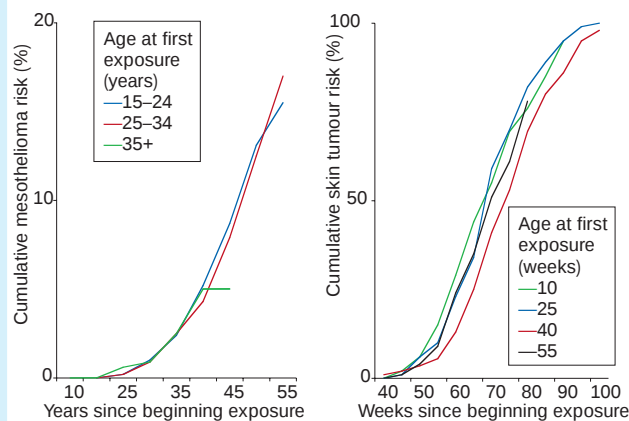
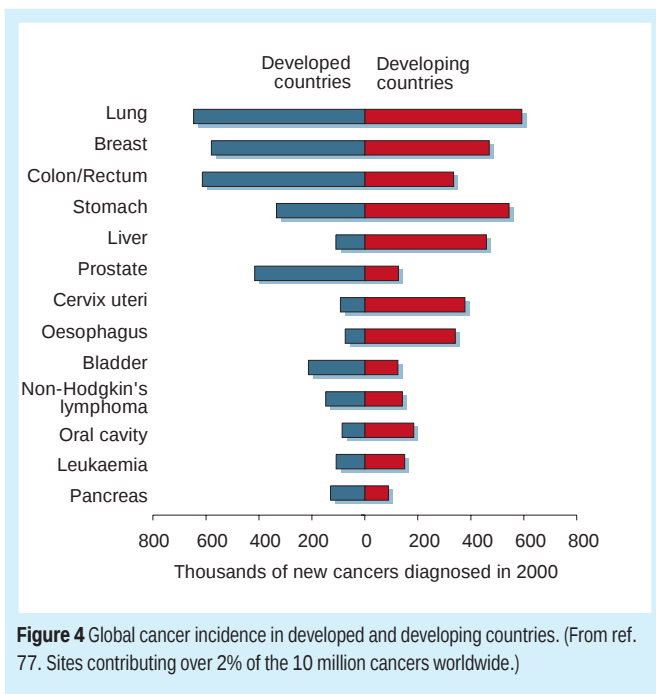


Figure 3 Age has no effect on susceptibility to some carcinogens. Left panel, cumulative mesothelioma risk in US insulation workers. Right panel, cumulative skin tumour risk in mice treated weekly with benzo(a)pyrene. Mesothelioma rates in humans⁶⁵ and skin tumour rates in mice⁶⁴ depend on time since first carcinogenic exposure but not on age, suggesting an initiating effect of these carcinogens. Lung cancer incidence in smokers depends on duration of smoking but not on age, and stops increasing when smoking stops⁶⁷, indicating both early- and late-stage effects. Radiation-induced cancer incidence increases with age at exposure above age 20, suggesting predominantly late-stage effects³, although the large effect of childhood irradiation also indicates an early-stage effect.



could be substantial, but their total contribution to cancer incidence will not be known until data on risk factors and extensive genotyping are available for very large numbers of patients and controls.

Familial risks for common cancers

Highly penetrant hereditary conditions such as polyposis coli, Li-Fraumeni syndrome and familial retinoblastoma cause at most a few per cent of the majority of cancers, and from an epidemiological perspective the genetic basis of the roughly twofold increase in incidence of the same type of cancer in first-degree relatives of patients with most common cancers (ref. 59 and article in this issue by Ponder, pages 336–341) is a more important question. A mendelian gene must confer a risk an order of magnitude greater than that in non-carriers for the risk in patients' relatives to be twice that in the general population⁶⁰. The individual effects of the common polymorphisms described above are thus far too small to account for much of this familial risk, although synergistic combinations could do so.

An important first step is to estimate the proportion of cancers of each type that arise in susceptible individuals and the contribution to this overall familial effect that can be accounted for by known genes. Breast cancer is so common that twin and sibling risks can be estimated fairly precisely. The high risk in patients' identical twins indicates that susceptible women contribute a high proportion, and perhaps even the majority, of overall breast cancer incidence⁶¹. This must be due mainly to 'low-penetrance' genes. Most multiple-case families⁶², but only about 2% of all cases⁶³, are due to mutations in *BRCA1* or *BRCA2*. The genetic epidemiology of colon cancer is quantitatively similar, although there has been less extensive sequencing of known genes in unselected cases. The quantitative contribution of penetrant genes to overall cancer incidence for common cancers such as prostate, melanoma and stomach has not yet been determined. 'Low-penetrance' susceptibility conferring a site-specific lifetime risk of the order of 30–50% may underlie many cancers, but it would almost never cause large numbers of cases in a family. If many genes contribute to the large genetic effects that seem to underlie many common cancers, they may be discoverable only through advances in our understanding of carcinogenic mechanisms.

Mechanisms of carcinogenesis

Age-incidence patterns for non-hormone-dependent carcinomas, and the effects of timing and dose-level of various agents alone and in combination (particularly smoking, alcohol, ionizing radiation and

some occupational carcinogens), are parsimoniously explained by the 'multi-stage' model of carcinogenesis. The evidence underlying this early work, which preceded the identification of any of the hypothesized sequence of heritable events in human carcinogenesis³, seems sometimes to be rather neglected. For example, the epidemiological and experimental evidence that somatic ageing processes per se play little or no role in carcinogenesis (Fig. 3 and refs 3, 64, 65) was not discussed in a recent review on cancer and ageing that argued exactly the opposite⁶⁶.

The incidence rate of cancer is presumably proportional both to the rate of the final rate-limiting step in carcinogenesis and to the number of premalignant cells that have undergone all but this final step. The rapid increase in the lung cancer incidence rate among continuing smokers ceases when they stop smoking, the rate remaining roughly constant for many years in ex-smokers⁶⁷. The fact that the rate does not fall abruptly when smoking stops indicates that the mysterious final event that triggers the clonal expansion of a fully malignant bronchial cell is unaffected by smoking, suggesting a mechanism involving signalling rather than mutagenesis. Such data are still generating new mechanistic hypotheses^{61,68}.

The future of cancer epidemiology

Over the next decade, cancer epidemiologists will be increasingly preoccupied with genetically susceptible subgroups. Comparison of the DNA in cancerous and normal cells from the same patient may lead directly to the identification of most of the genes that are commonly mutated in carcinogenesis. Candidate genes are also being identified on the basis of structural homologies from the human genome sequence. Extensive sequence or SNP comparisons between affected relatives and between cancer patients and controls may define combinations of polymorphisms or inherited defects in such genes that identify a few percent of the population whose average lifetime risk may be as high as 50% for a particular cancer. An alternative possibility is that susceptibility genes underlying phenotypic characteristics such as mammographic density⁶⁹ or chromosomal instability⁷⁰ that correlate with cancer risk and exhibit mendelian segregation will be found by linkage. Genes involved in DNA repair are likely to prove particularly important. Assays for defective DNA repair correlate consistently with substantially increased susceptibility⁷¹, and chromosomal aberrations predict increased cancer risk irrespective of carcinogenic exposure⁷².

Once they are identified, susceptible people might benefit disproportionately from screening or prophylaxis, while those at low risk would be reassured. But there will also be penalties. A different susceptible minority will be identified for each disease, and a high proportion of the population may eventually suffer the consequences of being classed as genetically susceptible to some major risk. The hazards of screening for cancer susceptibility are illustrated by the

Cancer site	No. of cases	Expected no.	Ratio
Non-melanoma skin	127	5.1	24.7
Thyroid and other endocrine	30	2.1	14.3
Mouth, tongue and lip	22	1.6	13.8
Cervix, vulva and vagina	39	3.6	10.8
Non-Hodgkin's lymphoma	25	2.4	10.3
Kidney and ureter	32	3.5	9.1
Bladder	26	4.7	5.5
Colorectal	38	10.5	3.6
Lung	30	12.5	2.4
Brain	10	4.1	2.4
Prostate	11	5.2	2.1
Melanoma	7	4.1	1.7
Breast	15	13.6	1.1

*Source: ref. 40.

widespread introduction of testing for prostate-specific antigen in the United States, which has reduced prostate cancer mortality only marginally but has led to a sharp increase in recorded incidence and considerable post-operative psychosexual and physical morbidity. Striking gene–environment interactions may be discovered, but most causes of cancer are likely to increase the risk by a smaller amount but a similar factor in those who are less susceptible. If smokers are less likely to stop smoking on discovering that their lifetime lung cancer risk is ‘only’ 10%, the population death rate might even be increased by such knowledge.

Advances in genetic and molecular understanding will increasingly enable epidemiologists to quantify the relationships between risk factors and specific events in carcinogenesis. Direct monitoring of changes in the genes that underlie carcinogenesis or their products is likely to provide sensitive and specific measures that can be correlated both with cancer incidence and with exposure to carcinogenic agents or activities. Characteristic mutations in DNA from subclinical cancers⁷³ or their precursor lesions^{74,75} can already be quantified, and serum levels of hormones such as oestrogen²³ and prolactin²⁴, or growth factors such as insulin-like growth factor-I, as well as chromosomal damage itself⁶⁷, are predictive of increased risk for certain cancers.

The most significant developments in cancer epidemiology may result from discoveries in virology and tumour immunology. The speculation that unidentified viruses (perhaps including some animal viruses⁷⁶) are associated with many human cancers is consistent with the large increase in overall cancer rates seen in immunosuppressed patients⁴⁰. The difficulty is that an unknown virus might mimic the epidemiological effects of dietary or genetic mechanisms. Thus, for example, the migrant patterns for prostate cancer (Fig. 1) might be due partly to an endemic infection, as they are for stomach cancer³⁰. Viruses usually act synergistically with other carcinogens and therefore provide alternative approaches to risk reduction. Perhaps the best way to prevent mesothelioma following heavy asbestos exposure will be by targeting SV-40 (ref. 38). The crucial issue is which of the increased risks in Table 1 reflect an unknown viral aetiology and which reflect immunosurveillance targeted at non-viral tumour markers. Some cancers may well be preventable by vaccination with tumour-specific antigens or by some less specific immunostimulation (see articles in this issue by Rosenberg, pages 380–384, and Appelbaum, pages 385–389).

Current priorities in cancer prevention

The large differences in the pattern of cancer incidence between developed and developing countries (Fig. 4)⁷⁷ imply different priorities for prevention, but at an individual level the most important difference is between smokers and non-smokers, particularly in developed countries. Table 2 shows approximate percentages of future cancer deaths in the United States that would be avoided by successively removing the effects of smoking, known infections, alcohol, sunlight, current occupational and environmental pollution, inactivity and obesity. The additional effect of specific dietary recommendations such as those of the American panel¹⁵ is much more speculative. Avoidance of overweight and prevention or treatment of oncogenic infections are the most important aims for non-smokers; but it is absurd for smokers in the West to worry about anything except stopping smoking.

Tobacco causes one-third of all cancer deaths in developed countries. About 15% of cancers worldwide are caused by known infectious agents³⁴. HBV alone causes almost as many cancers as smoking in China, and can be prevented by vaccination. HPV vaccines that are already being tested may be able to prevent almost all cervical cancers⁷⁸, and if the prevalence of *Helicobacter pylori* can be reduced, many stomach cancers would be avoided. The belated elimination of asbestos by many Western countries will eventually prevent the great majority of mesotheliomas and many lung cancers. (Whether almost all mesotheliomas are caused by crocidolite,

Table 2 US Cancer deaths that would be avoided by eliminating known risks

Cause	Deaths (%) avoided after removing preceding causes	
	Current smokers*	Non-smokers
Smoking	60	—
Known infections†	2	5
Alcohol‡	0.4	1
Sunlight	0.4	1
Air pollution§	0.4	1
Occupation	0.4	1
Lack of exercise¶	0.4	1
Diet#		
Overweight (BMI>25 kg m ⁻²)	4	10
Other dietary factors	4–12?	10–30?
Presently unavoidable☆	About a quarter	At least half

Approximate percentages of future cancer deaths among smokers and non-smokers in the United States that would be prevented by successive elimination of smoking, known infections, alcohol, sun exposure, current levels of workplace and air pollution, lack of exercise, and obesity. The estimates for the residual effect of specific dietary factors in people of normal weight are much less certain.

*For causes other than smoking, the table shows the additional reductions that could be achieved after smoking has been eliminated, so synergism with smoking is ignored. About 60% of cancers among smokers are due to smoking⁴, so all other percentages for smokers are 40% of those for non-smokers. The percentages for smokers relate to cancer risks during a smoker's lifetime and ignore risks during the 15 years of extra life gained on average when a smoking-related death is prevented by stopping smoking.

†All cervical cancers (HPV) plus half of all stomach cancers (*H. pylori*) gives 4% for non-smokers. A further 1% is added for all other infections. These two cancers and those caused by EBV, HBV and HCV are much more common in many developing countries. HBV alone, which can be prevented by vaccination, causes almost as many cancers as smoking in China. About 15% of cancers worldwide would be prevented if all known carcinogenic infections were eliminated³⁴.

‡Alcohol contributes about 0.5% due to upper aerodigestive cancers in non-smokers and a similar number due to breast cancer, although the reduction in heart disease among moderate drinkers outweighs their increased cancer risk.

§The lung cancer risks caused by radon and smoke from fossil fuels are largely confined to smokers. But lung cancer accounts for 2% of all cancer deaths in non-smokers, and a substantial proportion may be due to indoor and outdoor air pollution, including indoor radon. Mesotheliomas caused by non-occupational asbestos exposure contribute less than 0.1% of cancer mortality.

||Construction and maintenance workers are still sometimes exposed to asbestos and some workers are quite heavily exposed to dusts containing crystalline silica. Occupational exposure to chemicals causes some bladder and other cancers. Current exposures to asbestos and other known occupational carcinogens are much less than they were 40 years ago in most developed countries, and occupational lung cancer risks are much lower in non-smokers. Heavy exposure to occupational carcinogens still occurs in many developing countries.

¶As well as preventing heart disease, regular exercise reduces the risk of colon cancer and probably also of breast cancer, although the quantitative effect is uncertain^{10,15}.

#Mortality from various cancers and other causes increases progressively with increasing body-mass index (BMI) above 25 kg m⁻², and overall mortality also increases as BMI falls below about 22 kg m⁻² (refs 17, 18). The additional benefits of dietary recommendations for those with BMI between 22 and 25 kg m⁻² are very uncertain.

☆A few per cent of cancers are avoidable by the combined effects of measures that most people would not consider, such as having many children at an early age, followed by oophorectomy (surgical removal of one or both ovaries). Modern chemotherapy, radiotherapy and radiography usually do more good than harm, and cosmic radiation is unavoidable. The presumably environmental causes of the increased incidence of testicular cancer and non-Hodgkin's lymphoma are unavoidable because they are not yet known.

amosite or tremolite is still contentious^{79,80}, but all forms of asbestos cause lung cancer.) Various cancer screening tests are partially effective, and cervical screening is very effective.

The threat to epidemiology of the new ethics

Government action is essential to protect epidemiological research from the increasing burden of ‘ethical’ laws or conventions that bear no relation to patients’ physical or psychological well-being. Examples in Britain include the recent directive by the General Medical Council that doctors who notify their patients to cancer registries without obtaining fully informed consent may face disciplinary action, and the earlier delay in introducing anonymous HIV testing of discarded blood samples on the grounds that ‘screening must confer some benefit on the patient’⁸¹. Under the Data Protection Act it may even be illegal to use historical personnel records to study the

mortality of factory workers, as it is impractical to obtain informed consent from all former workers. Legislation is urgently needed to restore the long-established principle that consent is not mandatory for access to medical or civil records for bona fide medical research that has no effect on the individuals concerned and has been approved by a competent ethics committee. □

1. Doll, R. & Peto, R. The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *J. Natl Cancer Inst.* **66**, 1191–1308 (1981).
2. World Health Organization. *Prevention of Cancer (WHO Tech. Rep. 276)* (WHO, Geneva, 1964).
3. Doll, R. An epidemiological perspective of the biology of cancer. *Cancer Res.* **38**, 3573–3583 (1978).
4. Peto, R., Lopez, A. D., Boreham, J., Heath, C. & Thun, M. *Mortality from Tobacco in Developed Countries, 1950–2000* (Oxford Univ. Press, Oxford, 1994).
5. Doll, R. & Hill, A. B. Smoking and carcinoma of the lung. Preliminary report. *Br. Med. J.* **ii**, 739–748 (1950).
6. Wynder, E. L. & Graham, E. A. Tobacco smoking as a possible etiologic factor in bronchogenic carcinoma. *J. Am. Med. Assoc.* **143**, 329–336 (1950).
7. Wald, N., Doll, R. & Copeland, G. Trends in tar, nicotine, and carbon monoxide yields of UK cigarettes manufactured since 1934. *Br. Med. J. (Clin. Res. Ed.)* **282**, 763–765 (1981).
8. Peto, R. *et al.* Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *Br. Med. J.* **321**, 323–329 (2000).
9. Doll, R. Cancers weakly related to smoking. *Br. Med. Bull.* **52**, 35–49 (1996).
10. Doll, R. & Peto, R. in *Oxford Textbook of Medicine* 4th edn (eds Warrell, D. A., Cox, T. M. & Firth, J. D.) (Oxford Medical Publications, Oxford, in the press).
11. Liu, B. Q. *et al.* Emerging tobacco hazards in China: 1. Retrospective proportional mortality study of one million deaths. *Br. Med. J.* **317**, 1411–1422 (1998).
12. Peto, R., Lopez, A. D., Boreham, J., Thun, M. & Heath, C. Mortality from tobacco in developed countries: indirect estimation from national vital statistics. *Lancet* **339**, 1268–1278 (1992).
13. Working Group on Diet and Cancer of the Committee on Medical Aspects of Food and Nutrition Policy. *Nutritional Aspects of the Development of Cancer* (Dept Health Rep. Health Social Subjects 48) (The Stationery Office, London, 1998).
14. IARC. *Some Naturally Occurring Substances: Food Items and Constituents, Heterocyclic Aromatic Amines and Mycotoxins* (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 56) (IARC, Lyon, 1993).
15. American Institute for Cancer Research/World Cancer Research Fund. *Food, Nutrition and the Prevention of Cancer: A Global Perspective* (American Institute for Cancer Research, Washington, 1997).
16. Josefson, D. Obesity and inactivity fuel global cancer epidemic. *Br. Med. J.* **322**, 945 (2001).
17. Bergstrom, A., Pisani, P., Tenet, V., Wolk, A. & Adami, H. O. Overweight as an avoidable cause of cancer in Europe. *Int. J. Cancer* **91**, 421–430 (2001).
18. Calle, E. E., Thun, M. J., Petrelli, J. M., Rodriguez, C. & Heath, C. W. Body-mass index and mortality in a prospective cohort of U.S. adults. *N. Engl. J. Med.* **341**, 1097–1105 (1999).
19. Hennekens, C. H. *et al.* Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. *N. Engl. J. Med.* **334**, 1145–1149 (1996).
20. Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N. Engl. J. Med.* **330**, 1029–1035 (1994).
21. Omenn, G. S. *et al.* Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *N. Engl. J. Med.* **334**, 1150–1155 (1996).
22. Janne, P. A. & Mayer, R. J. Chemoprevention of colorectal cancer. *N. Engl. J. Med.* **342**, 1960–1968 (2000).
23. Hankinson, S. E. *et al.* Plasma sex steroid hormone levels and risk of breast cancer in postmenopausal women. *J. Natl Cancer Inst.* **90**, 1292–1299 (1998).
24. Hankinson, S. E. *et al.* Plasma prolactin levels and subsequent risk of breast cancer in postmenopausal women. *J. Natl Cancer Inst.* **91**, 629–634 (1999).
25. IARC. *Hormonal Contraception and Post-Menopausal Hormonal Therapy* (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 72) (IARC, Lyon, 1999).
26. Casagrande, J. T. *et al.* 'Incessant ovulation' and ovarian cancer. *Lancet* **2**, 170–173 (1979).
27. Pike, M. C., Krailo, M. D., Henderson, B. E., Casagrande, J. T. & Hoel, D. G. 'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer. *Nature* **303**, 767–770 (1983).
28. Newcomb, P. A. *et al.* Lactation and a reduced risk of premenopausal breast cancer. *N. Engl. J. Med.* **330**, 81–87 (1994).
29. Swerdlow, A. J. in *Principles and Practice of Genitourinary Oncology* (eds Raghavan, D., Scher, H. I., Leibel, S. A. & Lange, P.) 643–652 (Lippincott-Raven, New York, 1997).
30. IARC. *Schistosomes, Liver Flukes and Helicobacter pylori* (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 61) (IARC, Lyon, 1994).
31. Walboomers, J. M. *et al.* Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J. Pathol.* **189**, 12–19 (1999).
32. IARC. *Human Papillomaviruses* (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 64) (IARC, Lyon, 1995).
33. IARC. *Hepatitis Viruses* (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 59) (IARC, Lyon, 1994).
34. Parkin, D. M., Pisani, P., Munoz, N. & Ferlay, J. The global health burden of infection associated cancers. *Cancer Surv.* **33**, 5–33 (1999).
35. IARC. *Epstein-Barr Virus and Kaposi's Sarcoma Herpesvirus/Human Herpesvirus 8* (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 70) (IARC, Lyon, 1997).
36. IARC. *Human Immunodeficiency Viruses and Human T-cell Lymphotropic Viruses* (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 67) (IARC, Lyon, 1996).
37. Kinlen, L. J. Infection and childhood leukemia. *Cancer Causes Cont.* **9**, 237–239 (1998).
38. Carbone, M., Rizzo, P. & Pass, H. Simian virus 40: the link with human malignant mesothelioma is well established. *Anticancer Res.* **20**, 875–877 (2000).
39. Kinlen, L. J., Sheil, A. G., Peto, J. & Doll, R. Collaborative United Kingdom-Australasian study of cancer in patients treated with immunosuppressive drugs. *Br. Med. J.* **2**, 1461–1466 (1979).
40. Birkeland, S. A. *et al.* Cancer risk after renal transplantation in the Nordic countries, 1964–1986. *Int. J. Cancer* **60**, 183–189 (1995).
41. Peto, J., Decarli, A., La Vecchia, C., Levi, F. & Negri, E. The European mesothelioma epidemic. *Br. J. Cancer* **79**, 666–672 (1999).
42. Wagner, J. C., Sleggs, C. A. & Marchand, P. Diffuse pleural mesothelioma and asbestos exposure in the North Western Cape Province. *Br. J. Ind. Med.* **17**, 260–271 (1960).
43. Doll, R. Mortality from lung cancer in asbestos workers. *Br. J. Ind. Med.* **12**, 81–86 (1955).
44. Venitt, S. & Phillips, D. H. in *Environmental Mutagenesis* (eds Phillips, D. H. & Venitt, S.) 1–20 (Bios Scientific, Oxford, 1995).
45. Dockery, D. W. *et al.* An association between air pollution and mortality in six U.S. cities. *N. Engl. J. Med.* **329**, 1753–1759 (1993).
46. Hein, D. W. *et al.* Molecular genetics and epidemiology of the NAT1 and NAT2 acetylation polymorphisms. *Cancer Epidemiol. Biomarkers Prev.* **9**, 29–42 (2000).
47. Cartwright, R. A. *et al.* Role of N-acetyltransferase phenotypes in bladder carcinogenesis: a pharmacogenetic epidemiological approach to bladder cancer. *Lancet* **2**, 842–845 (1982).
48. Doll, R. *Possibilities for the Prevention of Cancer* (Royal Society Lecture for the Public, 13 November 1986) (Royal Society, London).
49. Johns, L. E. & Houlston, R. S. Glutathione S-transferase μ 1 (GSTM1) status and bladder cancer risk: a meta-analysis. *Mutagenesis* **15**, 399–404 (2000).
50. Houlston, R. S. Glutathione S-transferase M1 status and lung cancer risk: a meta-analysis. *Cancer Epidemiol. Biomarkers Prev.* **8**, 675–682 (1999).
51. Dunning, A. M. *et al.* A systematic review of genetic polymorphisms and breast cancer risk. *Cancer Epidemiol. Biomarkers Prev.* **8**, 843–854 (1999).
52. Woodage, T. *et al.* The APC1307K allele and cancer risk in a community-based study of Ashkenazi Jews. *Nature Genet.* **20**, 62–65 (1998).
53. Krontiris, T. G., Devlin, B., Karp, D. D., Robert, N. J. & Risch, N. An association between the risk of cancer and mutations in the HRAS1 minisatellite locus. *N. Engl. J. Med.* **329**, 517–523 (1993).
54. Bennett, W. P. *et al.* Environmental tobacco smoke, genetic susceptibility, and risk of lung cancer in never-smoking women. *J. Natl Cancer Inst.* **91**, 2009–2014 (1999).
55. Ambrosone, C. B. *et al.* Cigarette smoking, N-acetyltransferase 2 genetic polymorphisms, and breast cancer risk. *J. Am. Med. Assoc.* **276**, 1494–1501 (1996).
56. Weinberg, C. R. & Sandler, D. P. Gene-by-environment interaction for passive smoking and glutathione S-transferase M1? *J. Natl Cancer Inst.* **91**, 1985–1986 (1999).
57. IARC. *Tobacco Smoking* (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 38) (IARC, Lyon, 1986).
58. Ma, J. *et al.* Methylene tetrahydrofolate reductase polymorphism, dietary interactions, and risk of colorectal cancer. *Cancer Res.* **57**, 1098–1102 (1997).
59. Houlston, R. S. & Peto, J. in *Genetic Predisposition to Cancer* (eds Eeles, R. A., Ponder, B. A. J., Easton, D. F. & Horwich, A.) 208–226 (Chapman & Hall, London, 1996).
60. Peto, J. in *Cancer Incidence in Defined Populations. Banbury Report 4* (eds Cairns, J., Lyon, J. L. & Skolnick, M.) 203–213 (Cold Spring Harbor Laboratory Press, 1980).
61. Peto, J. & Mack, T. M. High constant incidence in twins and other relatives of women with breast cancer. *Nature Genet.* **26**, 411–414 (2000).
62. Ford, D. *et al.* Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* **62**, 676–689 (1998).
63. Peto, J. *et al.* Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. *J. Natl Cancer Inst.* **91**, 943–949 (1999).
64. Peto, R., Parish, S. E. & Gray, R. G. in *Age-Related Factors in Carcinogenesis* (IARC Sci. Publ. 58) (eds Likhachev, A., Anisimov, V. & Montesano, R.) 43–53 (IARC, Lyon, 1986).
65. Peto, J. in *Models, Mechanisms and Aetiology of Tumour Promotion* (IARC Sci. Publ. 56) (eds Borzsonyi, M., Day, N. E., Lapis, K. & Yamasaki, H.) 359–371 (IARC, Lyon, 1984).
66. DePinho, R. A. The age of cancer. *Nature* **408**, 248–254 (2000).
67. Halpern, M. T., Gillespie, B. W. & Warner, K. E. Patterns of absolute risk of lung cancer mortality in former smokers. *J. Natl Cancer Inst.* **85**, 457–464 (1993).
68. Janin, N. A simple model for carcinogenesis of colorectal cancers with microsatellite instability. *Adv. Cancer Res.* **77**, 189–221 (2000).
69. Pankow, J. S. *et al.* Genetic analysis of mammographic breast density in adult women: evidence of a gene effect. *J. Natl Cancer Inst.* **89**, 549–556 (1997).
70. Roberts, S. A. *et al.* Heritability of cellular radiosensitivity: a marker of low-penetrance predisposition genes in breast cancer? *Am. J. Hum. Genet.* **65**, 784–794 (1999).
71. Berwick, M. & Vineis, P. Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review. *J. Natl Cancer Inst.* **92**, 874–897 (2000).
72. Bonassi, S. *et al.* Chromosomal aberrations in lymphocytes predict human cancer independently of exposure to carcinogens. European Study Group on Cytogenetic Biomarkers and Health. *Cancer Res.* **60**, 1619–1625 (2000).
73. Anker, P., Mulcahy, H., Chen, X. Q. & Stroun, M. Detection of circulating tumour DNA in the blood (plasma/serum) of cancer patients. *Cancer Metastasis Rev.* **18**, 65–73 (1999).
74. Ponten, F. *et al.* Molecular pathology in basal cell cancer with p53 as a genetic marker. *Oncogene* **15**, 1059–1067 (1997).
75. Khrapko, K. *et al.* Constant denaturing capillary electrophoresis (CDCE): a high resolution approach to mutational analysis. *Nucleic Acids Res.* **22**, 364–369 (1994).
76. zur Hausen, H. Proliferation-inducing viruses in non-permissive systems as possible causes of human cancers. *Lancet* **357**, 381–384 (2001).
77. Parkin, D. M., Bray, F. I. & Devesa, S. S. Cancer burden in the year 2000. The global picture. *Eur. J. Cancer* (special supplement; in the press).
78. Schiller, J. T. & Lowy, D. R. Papillomavirus-Like Particle Vaccines. *J. Natl Cancer Inst. Monogr.* **28**, 50–54 (2000).
79. Cullen, M. R. Chrysotile asbestos: enough is enough. *Lancet* **351**, 1377–1378 (1998).
80. McDonald, A. D. *et al.* Mesothelioma in Quebec chrysotile miners and millers: epidemiology and aetiology. *Ann. Occup. Hyg.* **41**, 707–719 (1997).
81. Black, D. *et al.* HIV testing on all pregnant women. *Lancet* **2**, 1277 (1987).

Acknowledgements

Parts of the section on environmental and lifestyle factors are based on the review on avoidable causes of cancer by Richard Doll and my brother Richard Peto¹ and their updated summary¹⁰. I am also grateful to them and to many other colleagues for commenting on earlier drafts, and to the Cancer Research Campaign for support.

AstraZeneca and Cancer

discovery from a global perspective

AstraZeneca is delighted to support this Nature Insight on cancer at such an exciting time to be in cancer research. In the last 20 years, many new molecular targets have been identified from basic research and these are now leading to new agents for the treatment of cancer. The first publication of the human genome sequence has been completed and the race continues apace to find the best ways of using these genomic data to the benefit of future cancer research efforts.

Significant successes in treating cancer have been achieved over the last 25 years such that some types of cancer (e.g. Hodgkin's, Burkitt's lymphoma, testicular) now have very high response rates, representing a 50-60% decrease in death rate. Also, the anti-hormonal treatments of breast and prostate cancers have significantly increased the disease-free survival and quality of life for cancer patients. Nevertheless, cancer remains a disease of high unmet clinical need where life expectancy can often be short. In 1999, there were over 12 million new cases of cancer diagnosed and 7 million deaths. Cancer is the leading cause of death in Japan and is predicted to be the leading cause of death in the US by 2005. These raw statistics illustrate some of the elements of the scale of the patient need. In addition, the nature of the disease as we know it, through personal experience and through family and friends, is a strong motivator to continuing to persevere with cancer research.

For such a complex disease, it is essential that we tackle it with diverse thinking and practice. The Insight articles illustrate very succinctly the multi-faceted research being done to address this important disease. Our collective research findings and developments are set to change the range of cancer therapies available to patients over the next 10-20 years.

In the future, we hope that current agents will be improved upon but more importantly that the truly novel approaches, currently in pre-clinical and clinical testing, really will make a difference to cancer patients. It is going to become increasingly important to treat the patient as well as trying to kill or cure the tumour. Agents with enhanced selectivity for stopping the progression of cancer will be found and these agents will be used earlier in disease to prevent cancers from occurring. Good progress is being made and the pace is quickening. For example, less than a decade ago the field of angiogenesis was in its infancy and endothelial cell growth factor signalling was poorly characterised. Today, angiogenesis is one of the largest areas of focus for academic research and drug discovery within the Pharmaceutical Industry. Enticingly, anti-angiogenesis agents (e.g. specifically targeted at the endothelial growth factors) are now poised to benefit cancer patients as the novel agents progress rapidly through clinical trials and blockers of the EGF signalling pathway are producing very exciting clinical benefits.

With this view of the past and present, the Pharmaceutical Industry has a mission to develop new cancer therapies. AstraZeneca is committed to maintaining, through work at the two major cancer research sites in Boston, US and Macclesfield, UK, its leading position in delivering drugs for the treatment of cancer. In addition, AstraZeneca supports basic scientific research in cancer through a global network of research and clinical collaborations.

The recent advances in thinking and practise within cancer research, described in this publication, capture the challenges that lie ahead. In the rapidly progressing era of genomic, high throughput, informatics-based research that we are now in, we must continue to harness and exploit the new and apply it to our current understanding of cancer as a disease, to the benefit of cancer patients world-wide.

Dr Les Hughes

VP Global Cancer and Infection Research, AstraZeneca.

AstraZeneca 

Cultural revolutions

New ideas for the microbiologist include BugStoppers and bug counters.

Lid_{Bac}

Eppendorf www.eppendorf.com

Tailor-made for small-scale cultivation

Lid_{Bac} lids reduce excess medium and plastic waste by allowing miniprep bacteria cultivation in the test tube. The polypropylene membrane lid includes a gas-permeable, hydrophobic, bacteria-proof membrane with a pore size of 0.2 µm. After the culture has been vaccinated, the membrane lid is pressed onto the test tube. The culture is then available for plasmid preparation. The lids can be used in all processes in which aeration, drying or similar action is required, with no need to transfer the sample to a new container. They come supplied with 2.0 ml Safe-Lock micro test tubes, and can also be used with 1.5 and 2.0 ml Eppendorf tubes.

MBS 1 filtration system

Schleicher & Schuell www.s-und-s.de

Testing times for microbiological control

Consisting of funnel, apparatus and membrane, the MBS 1 filtration system uses an autoclavable plastic funnel that eliminates the need for flame-scarfing associated with sterilization of traditional stainless steel equipment and achieves a time saving of approximately 30 seconds per test. A large funnel capacity accommodates foaming liquids and there is a safe-sealing mechanism.

BioCoat dishes

BD Biosciences www.bd.com

Cover-slip bottom dishes

In these 35 mm poly-D-lysine dishes, the cover-slip floor is made of optically clear, non-neurotoxic glass with a low background fluorescence. Suggested applications include high-resolution, inverted, phase-contrast and confocal

microscopy, fluorescence imaging in live cells and micro-manipulations.

Steam sterilizer

Hotpack www.hotpack.com

Sterilization for the research lab

The Hotpack sterilizer has a load probe and Fourier number program to facilitate precise liquid sterilization regardless of load size and to reduce the risk of 'boil-over' or media degradation. The unit features a PLC-based controller with colour touch-screen permitting ±5°C temperature uniformity, four liquid, four pre-vacuum, and one gravity program, an automatic leak test and adjustable key parameters for customizing cycles. Supplied in a range of chamber sizes and configurations, with standard vertical or horizontal power doors, the sterilizer includes two standard chamber penetrations for additional thermocouple or load probes for mapping or product qualification testing. All vessels are designed for a working pressure of 45 psig and are supplied with an ASME form U-1. A range of optional extras is available.



PETG diagnostic bottles

Nalgene www.nalgenunc.com

A new line in sterile bottles

Supplied in tray packs in 5, 10 or 20 ml capacities, PETG diagnostic bottles are single-use, pre-sterilized systems ready for use. Complementing the company's PETG square media bottles, the diagnostic bottles are visually clear and suitable for use as storage and shipping containers for small-volume reagents in both laboratory and packaging applications. The bottles and 20 mm coloured lined closures have been tested for pyrogens using LAL gel clot methods and are sterile to 10⁻⁶ SAL.

HHV-8 ELISA kit

Advanced Biotechnologies www.abionline.com

Explicit expression

This new antibody test kit for detection of latent IgG antibody to human herpesvirus-8 (HHV-8; Kaposi's sarcoma herpesvirus) in serum/plasma is made with purified full-length recombinant HHV-8 latent protein, known as ORF 73 (LNA-1). ORF 73 is highly immunogenic and, since it is expressed in a baculovirus system, is free of the nonspecific protein reactivity often found with recombinant proteins expressed in *E. coli* cell systems.

Cytochrome c ELISA kit

Bender MedSystems

www.bendermedsystems.com

Ready reckoner for cell apoptosis

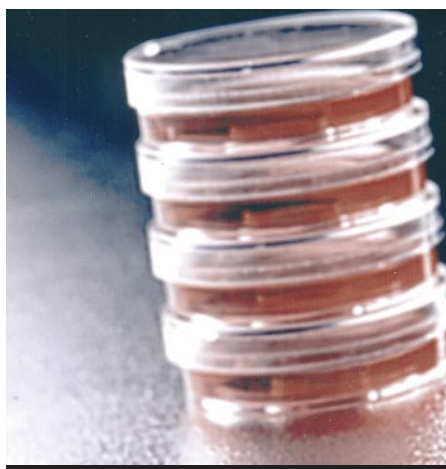
Benders new ready-to-use assay is designed for quantitative detection of human cytochrome c in cell culture lysates, human whole blood or serum. Cytochrome c indicates activation of apoptotic programs in cells and recent research involving serum from patients with haematological malignancies has suggested that serum cytochrome c might also serve as a marker for onset of apoptosis and cell turnover *in vivo*.

ProtoCOL RGB

Synbiosis www.synbiosis.com

All coloured colonies counted

The latest enhanced version of the ProtoCOL RGB microbiological colony counter automatically differentiates and counts coloured colonies using a new detection algorithm. All colony types of all colours can be differentiated on any colour of agar. A live digital image is displayed, which can then be saved and printed as a full-colour image. Specially designed illumination and camera control via image analysis software automatically compensates for varia-



Optical glass is the focus of BD's dish — and controlling steam the USP for the Hotpack sterilizer.

tions between plates or samples. Suggested uses include distinguishing *E. coli* from coliforms, determining antibiotic potency, clinical identification of pathogens and genetic screening in yeasts and bacteria.

Bioplus 2002

Technovation Systems www.cleanroomSys.com

Top of the class

This low-noise, fan-powered duct filter uses half the power of the standard model 3001B filter. Each 2002 is used with four terminal HEPA's in series for Class 0.1–100 rooms to achieve an overall filtration efficiency of 99.999999+%. No terminal HEPA's are necessary for clean-rooms of Class 1000 and above. The 2002 filter is bactericidal and inhibits bacterial growth on the filters, minimizing the airborne bio burden.

BioFlo 110

New Brunswick Scientific www.nbsc.com

New research fermentor and bioreactors

This latest NBS catalogue of equipment for growing, detecting and storing cultures features the BioFlo 110, a new research fermentor/bioreactor that can control up to four vessels. The catalogue also includes a line of standard and custom-designed cGMP-compliant fermentors and cell culture systems, laboratory shakers, automated media preparation equipment and combinatorial chemistry shakers.

Classic

Pierre Guerin www.devicelink.com

Pilot and laboratory bioreactors

Designed for bacterial or cell cultures, the Classic range of bioreactors can be cleaned and sterilized on site. Fitted with an operator interface under Windows, the laboratory series is available in capacities of 10–30 litres, while the pilot series offers capacities of 40–275 litres. The open-frame units are pre-mounted and pre-wired, and are supplied with process optimization software modules.

MIDAScan AT

Trevigen www.trevigen.com

Mutation detection system

This scanning system utilizes DNA mismatch repair enzymes to identify the type and location of mutations within DNA sequences up to 1kb long without requiring DNA sequencing. It can be used to identify both T and A mutations. The control kit includes two DNA sequences that, when heated, mixed and cooled, form all possible DNA mismatches at 20-bp intervals. Labelled DNA can be used as a control for most mutation scanning methods. Because the mismatches are evenly spaced, the pattern of cleavage easily identifies the mismatch and generates a DNA size ladder. The control kits can also be used to investigate the activities of mismatch repair enzymes.



BugStopper: does what it says.

BugStopper

Whatman www.whatman.com

Plug away with this device

Designed for culture vessels that can be autoclaved and re-used, the BugStopper sterile venting closure is constructed from biosafe silicone and can withstand repeated autoclaving with no reduction in airflow or retention capability. It restricts bacteria and viruses from entering or exiting a vessel while allowing air/gas to pass through the membrane, eliminating the need for cotton plugs. The closure has a filter rating of 99.9% efficiency for both bacteria and viruses.

MicrobeLynx

Micromass www.micromass.co.uk

Fingerprint bacterial ID

The networked M@LDI mass analyser coupled with MicrobeLynx bioinformatics technology form the basis of this automated bacterial 'mass-fingerprinting' approach to speciation and typing of microorganisms. Macromolecules expressed on the surface of bacteria are sampled and characterized by molecular weight. The resulting mass spectrum provides a unique fingerprint for the species tested. Bacterial mass fingerprints of unknowns can be reliably matched against a database of quality-

controlled reference spectra. Suggested applications include medical microbiology, and the food, water, pharmaceutical and biotechnology industries.

Swiftlock

Astell Scientific www.astell.com

An easy opening for a compact autoclave

This range of compact 23-, 35- and 50-litre autoclaves features a 'super-quick' door mechanism for ease of access. A microprocessor control system ensures repeatable, safe and accurate sterilization and offers four pre-set programs for typical loads, with a 16-character, two-line LCD display to record cycle progress. Safety features include overheat protection, audible and visual alarms, pressure interlocks on doors and cooling locks for fuel cycles. An optional integral data printer records temperature, pressure and elapsed time for each cycle, together with cycle stage messages and number, thus providing a traceable gas record. Other options include load-sensed process timing, assisted cooling and a non-recirculation system for situations in which blockage or contamination may occur.

OncoQuick

Hexal Gentech www.hexal-gentech.com

Tumour cell enrichment

There is increasing clinical interest in detecting circulating tumour cells in bone marrow and whole blood and integrating the monitoring of minimal residual disease (MRD) in tumour staging and prognosis. OncoQuick, developed by Greiner Bio-One Ltd and Hexal Gentech, is designed to enrich circulating tumour cells from up to 25ml of whole blood. It uses a liquid density medium optimized for the specific enrichment of circulating tumour cells based on their buoyant density. The entire enrichment procedure can be completed within 45 minutes, yielding more than 10⁴ total mononuclear cells.

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

Polyclonal Antibodies & Custom Peptides

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

ResGen™
Invitrogen Corporation

U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our web site at
www.resgen.com

Proteasome inhibition?

~ epoxomicin ~
~ PR39 ~ lactacystin ~
~ clasto-lactacystin ~ β ~
~ lactone ~ P131 ~
~ MG-132 ~ MG-262 ~

www.proteasome.com

AFFINITY Research Products Ltd.
Mamhead Castle, Exeter, EX6 8HD, U.K.
T.+44(0) 1626 891010 – F.+44(0) 1626 891090
– Email info@proteasome.com –

Contacts

Publisher: Fabien Savenay

Editor: Paul Smaglik

Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

Group European Manager:

Leonie Welss (4954)

European Manager:

Nevin Bayoumi (4978)

European Sales Consultant:

Jonathan Bull (4973)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),

Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Sille Opstrup (4994)

Production Manager:

Billie Franklin
To send films and materials use
London address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: classified@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau

Tel + 33 (0) 1 43 20 16 51

Fax + 33 (0) 1 43 20 51 52

e-mail: c.nioxchateau@nature.com

Spain/ Portugal:

Amanda Wren

Tel + 34 91 3913835

Fax + 34 91 319 0091

e-mail: a.wren@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mails: p.phelan@nature.com

k.turner@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: classified@nature.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Corissa Salzman

Japan Head Office, Tokyo

Shin-Mitsuke Building (4F),

3-6 Ichigaya Tamachi Shinjuku-Ku,

Tokyo 162

Tel +81 3 3267 8751

Fax +81 3 3267 8746

e-mail: k.cowan@nature.com

Japan Manager:

Kate Cowan

Panning for gold

In the nineteenth century, fortune seekers headed to California to pan for gold. In the late twentieth century, hopeful scientists flocked there to prospect for proteins. Since then, other parts of the world have joined the biotech rush — often looking to the San Francisco Bay Area for cues about where and how to dig for new compounds and build new companies. But the original blueprint that the site of biotech's genesis provided to the world has changed dramatically — and with it the dynamics of employment there (see page 4).

For example, when Genentech helped to launch the biotech era 25 years ago, the company was relatively small and it staked its claim with one putative product. Now, the company — with its thousands of employees, multiple therapies on the market and several in the pipeline — more closely resembles a pharmaceutical giant than a start-up. The same is true for other more established companies such as Chiron.

Further blurring the distinction between biotech and pharmaceutical companies, many of the early northern California biotechs are owned at least partially by pharmaceutical concerns. Many of the newer ones were formed as the result of a partnership with, or as spin-offs from, other companies.

Within this protean landscape, the future of California biotech is being reshaped yet again. Several major investments in new university programmes and buildings, which will probably trigger another round of start-ups, spin-offs and tie-ins, look set to boost the employment market in an already crowded, yet prosperous, region. Other regions of the world looking to the Bay Area biotech model may want to ask themselves which Bay Area they're basing their plans upon: the one that was, that is, or is to come.

Paul Smaglik

Naturejobs editor



Contents

REGIONS

New investment reinvigorates the birthplace of biotech p4

MOVERS

NASA's Mars man returns to Ames, mathematician broadens horizons yet further, and more
Back page

WWW.NATURE.COM/ NATUREJOBS

Career centre
Information on the scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



LENI

The birthplace of biotech California

Davis sows seeds for future

‘Warm at the centre, cooling at the edges’ could sum up bioscience today in the San Francisco Bay Area. Queasy financial markets mean it is not a propitious time to take a new company public, and job hunters in support areas such as public relations, marketing or technical writing may have to look harder for work. But scientists with technical skills are more likely to find the Bay Area’s power blackouts, gridlocked freeways and ultra-costly housing to be the annoying costs of success rather than barriers to future advancement. Indeed, by many measures the fundamental factors for good job prospects are stronger than ever.

Mina Bissell, director of the Life Sciences Division at Lawrence Berkeley Lab, points to a combination in Northern California that few places in the world can match: four research universities (three with medical schools), two government labs, an entrepreneurial culture and venture capital community, plus its clusters of bioscience companies. The Bay Area, of course, is where biotech began; the industry’s progenitor, Genentech, in South San Francisco, celebrated its 25th anniversary this year. Today the region is home to 697 biomedical firms, including device companies, with another 100 in the Davis/Sacramento area, 70 miles away (see “Davis sows seeds for future”; a complete list of companies in the region is available at <http://www.chi.org/biomedicalresource.php>).

Companies range in size from giants such as Genentech, with 3,100 employees, to boutique start-ups such as Silicon Genetics in San Carlos, 10 miles south of San Francisco. Growth in the regional industry continues unabated, providing 85,000 direct jobs and generating \$8 billion in annual revenues. For example,

When researcher Dave Gilchrist noticed similarities in a key metabolic pathway shared by species in both the plant and animal kingdoms, he came up with a baculovirus-mediated method for activating programmed cell death. Gilchrist, based at the University of California at Davis, California’s premier agricultural university, hopes a biologically based method for preventing fungal rot disease in tomatoes will emerge from the work.

Informatics drives this sort of discovery in plant, animal and environmental

biotechnology, says Martina McGloughlin, director of both the biotechnology and life sciences informatics programmes at Davis. McGloughlin coordinates her university’s efforts with those of the Joint Genome Institute and Lawrence Livermore National Lab. Growth at Davis, if not quite as feverish as in the Bay Area, is nonetheless robust.

The campus is projected to grow by 5,000 students over the next 10 years, and faculty staff, particularly in genomics, proteomics and bioinformatics, are being recruited.

P. W.

LENI



The concentration of first-class facilities and scientists in the Bay Area attracts the best people, according to Mina Bissell.

South San Francisco is home to 80 biotech companies — many of them spin-offs from Genentech — up from 34 just two years ago. And university campuses in the area are undergoing a construction boom as well, which will result in more jobs in the future (see “Bay Area building boom”). The region’s appeal is evident in the funding numbers. At \$703 million in 2000, Northern California ranks only behind Boston and New York in National Institutes of Health (NIH) extramural grant money received. Eight times more NIH money per capita than the national average flows into the area.

The University of California at San Francisco (UCSF), the area's crown jewel biomedical research university, is the fourth-largest institutional recipient of NIH money, with \$266 million last year. Stanford, 30 miles to the south, came in at number nine. But it is in private funding that the Bay Area, birthplace of the venture-capital-funded start-up, outshines all other regions. According to the National Venture Capital Association, an industry group in Alexandria, Virginia, biotech companies here got over \$1 billion in 2000, with another \$1 billion for the medical/health sector. Across all sectors, the region was the destination for \$33.5 billion, up 72% from the previous year and nearly triple that of Greater New York, the next-closest part of the country.

Zach Hall, UCSF's vice-chancellor for research, adds, "We're already recruiting very high-quality talent to UCSF, and with our new campus, we'll be the most exciting and vibrant place in biomedical research". That research is done in ways that could not have been foreseen even a few years ago. The success of Genentech in cloning the genes for insulin and growth hormone in 1978-82, and turning these molecules into commercial products, encouraged ventures such as Tularik, Sugen, ICOS and many others.

Although the Genentech model extended the reach of science beyond what could be initiated and pursued in conventional academic and pharmaceutical settings, the trend of what was to come was not easy to discern in the early days. Then, conditioned by the classic model of the big pharmaceutical companies, newcomers to biotech aspired to contain their own R&D production, manufacturing, sales and marketing units. But of the early industry leaders, only Amgen remains a free-standing company, with 49% of Chiron owned by Novartis and 60% of Genentech by Roche.

"Today there isn't a single paradigm that works for every successful company," says David Gollagher, president of the California Healthcare Institute, an industry-supported think-tank in La Jolla. For example, COR Therapeutics, in South San Francisco, which makes the clot-busting drug Integrilin for angina and for use with angioplasty, has a marketing arrangement with Schering-Plough that gives their product international distribution while allowing COR to keep staff below 500. Other companies might contract out manufacturing for now to a big player such as Genentech, while gradually building up their own capacity. This sort of leveraging frees up company resources for R&D to develop more lead products now.

Bay Area building boom

With companies eager to reap bonanzas from emerging research, and universities keen to increase their prestige, a major building boom is under way in the Bay Area. The most ambitious project is University of California at San Francisco's (UCSF's) new Mission Bay campus, which will house 9,000 basic researchers plus a community of biotech companies.

Total staffing levels at UCSF are apt to rise by a third over the coming decade, according to Zach Hall, UCSF's vice-chancellor for research. Basic research at Mission Bay will be augmented by the lead unit of California's new \$200 million Institute for Bioengineering, Biotechnology and Quantitative Biomedicine, devoted to near-to-market research in such areas as instrumentation, high-

throughput screening, drug discovery, robotics, nanotechnology, molecular motors and proteomics.

Across the Bay, the University of California at Berkeley will pursue post-genome bioscience in a \$300-million complex of new buildings. And at Stanford a \$150-million structure will house 50 faculty staff and 400 students in biomedical research. Basic, applied and clinical science will converge so that traditionally separate disciplines can come together on research problems.

P. W.



Part of the UCSF 300-acre Mission Bay campus under construction at a disused industrial area.

As companies increase their partnerships with each other and with the academic sector, new problems inevitably crop up. Ways of dealing with conflicts of interest, or clashes between academic traditions of open inquiry and companies' focus on proprietary value will have to be devised.

But there is no going back to the old separation between public and private, says Chris Scott, assistant vice-chancellor for research at UCSF. "Biology these days is driven much more critically by the biotech industry as a source of innovation than it was even five years ago." Fluid, multidisciplinary collaborations, such as those at Berkeley, Mission Bay and Stanford (see "Bay Area building boom"), increasingly displace the department-centred research of the not-so-distant past.

Today, the promising new research direction is apt to originate from a chance encounter in the hallway, or a comment serendipitously overheard at lunch, rather than from the dean's directive. Part of the key to fostering this productive spontaneity is putting people close enough together to make contact. "People get sparks off each other," as UCSF's Zach Hall puts it. For many of today's best bioscientists, that electricity means the San Francisco Bay Area.

Potter Wickware is a science writer in San Francisco.



On a mission: Zach Hall (left) and Chris Scott review plans for the exciting and vibrant new campus at UCSF.

COMPUTER SCIENCE

Andrew Odlyzko, a mathematician and computer scientist known for his stance favouring free access to electronically published research (see *Nature's* online debate <http://www.nature.com/nature/debates/e-access>), is leaving AT&T in August to head up the University of Minnesota's Digital Technology Center. Odlyzko, who fondly calls his stint at AT&T "a 26-year postdoc", spent the last six years there leading the mathematics and cryptography research department.

While at AT&T, his work ranged from projects in security, distributed and parallel computing and, more recently, the economics of e-commerce and electronic publishing. Those more recent interests have prompted him to acquiesce to a headhunter's call soliciting him for the Minnesota position. "My research horizons have been getting broader," he says. "I have been getting interested in scientific policy and copyrights among other things."

At the university, where he will also serve as assistant vice-president for research, Odlyzko will supervise 150 technical people — most of whom will be postdocs, graduate students and visiting fellows. The centre has 14 faculty positions, half of which have been filled. The centre will house the Minnesota Supercomputing Institute, the Laboratory for Computational Science and Engineering, and the Telecommunications and Advanced Networking Laboratory. Its research focuses include telecommunications and advanced networking, software storage and Internet technologies, data storage and visualization, electronic commerce and digital publishing, bioinformatics and computational biology, among others.

at the way the company characterized some of those findings. He now is chief scientific officer at ValiGen, a European-US biotechnology company. Ravel, who spent 17 years with Servier, has developed drugs for both obesity and type 2 diabetes. "For me it's not really a big change," he says of the new position. However, at Servier, he focused more on small molecules, whereas at the biotech company he will spend more time working on proteins and peptides.

David Briscoe now holds the newly created position of commercial director for Pharmagene. Briscoe comes to the Cambridge-based company from Incyte Genomics, where he managed European operations for the Palo Alto, California-based company. He has 25 years' experience with life science companies, including stints with Chiron, Schering Plough and Bristol Myers Squibb.

SPACE SCIENCE

G. Scott Hubbard, who reorganized the NASA's Mars Program after losses of two robots last year, has relinquished the position. Hubbard launched the 2001 Mars Odyssey Program this spring. He has returned to NASA's Ames Research Center in Moffett Field, California. **Orlando Figueroa**, currently the deputy chief engineer for systems engineering at NASA Headquarters, Washington DC, replaced Hubbard as acting director of the programme. Figueroa spent 22 years of his career at the NASA Goddard Space Flight Center in Greenbelt, Maryland.



Andrew Odlyzko



G. Scott Hubbard

BIOTECH

Denis Ravel, former head of the Metabolism Department of the French pharmaceutical company Servier, has joined Genset in its Paris offices as Director of Pharmaceutical Development. Ravel fills some of the void created when **Bernard Bihain**, one of the company's two scientific directors, unexpectedly resigned in February, over what some biotech analysts have said was disagreement arising from Genset's transition from a gene research company to a drug-development firm. At the heart of the dispute is an anti-obesity drug, Famoxin, a protein that had shown some promise in preclinical animal experiments. Bihain reportedly was upset

GENOMICS

David Millhorn, chairman of the department of molecular and cellular physiology at the University of Cincinnati College of Medicine, will serve as director of the university's new Genome Research Institute. The institute will use information from the Human Genome Project to characterize genetic traits in various diseases. The institute will be located in a 360,000-square-foot laboratory facility donated by Aventis Pharmaceuticals. The university is currently recruiting other University of Cincinnati scientists and outside researchers to join the institute. Within five years, the university anticipates having a group of 40 to 50 principal investigators at the institute.



David Millhorn

CONTACTING US AT MOVERS

Please send any information on new appointments or job moves to: naturejobseditor@naturedc.com.